

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

© Copyright

by

Francia Jimenez

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 Alberto 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

| | |
|---|-------------|
| Acknowledgments | v |
| List of Tables | x |
| List of Figures | xiii |
| Acronyms | xxii |
| Abstract | xxiv |
| Publications and Outcomes | xxvi |
| 1 Introduction | 1 |
| 1.1 Data Science | 1 |
| 1.1.1 Feature Selection | 2 |
| 1.2 Motivation | 3 |
| 1.2.1 Multi-objective Optimization | 3 |
| 1.3 Goals | 4 |
| 1.4 Contributions of This Thesis | 5 |
| 1.5 Thesis Overview | 5 |
| 2 Feature Selection | 7 |
| 2.1 Feature Selection | 7 |
| 2.1.1 Classification of feature selection algorithms | 8 |
| 2.1.2 Definition of k -Feature Set Problem | 9 |
| 2.2 The (α, β) - k -Feature Set Problem | 11 |
| 2.2.1 Graph Representation for the (α, β) - k -Feature Set Problem | 12 |
| 2.2.2 Integer Programming Model of the (α, β) - k -Feature Set Problem | 13 |
| 2.3 Related Work with the (α, β) - k -Feature Set Problem | 15 |
| 2.4 Summary | 17 |
| 3 Multi-objective Optimization | 18 |
| 3.1 Multi-objective optimization | 18 |
| 3.1.1 Multi-objective optimization problem definition | 19 |
| 3.2 Approaches to solve multi-objective optimization problems | 20 |
| 3.2.1 Single-objective Reductions for MOP | 20 |

| | | |
|----------|---|-----------|
| 3.2.2 | Pareto Front-based Approaches for MOP | 22 |
| 3.3 | Multi-objective Performance Metrics | 23 |
| 3.3.1 | Hypervolume | 24 |
| 3.3.2 | C-measure | 24 |
| 3.4 | Optimization Approaches for Feature Selection | 25 |
| 3.5 | Multi-objective (α, β) - k -Feature Set Problem | 26 |
| 3.6 | Summary | 29 |
| 4 | The Multi-objective Memetic Algorithm for the (α, β)-k-Feature Set Problem | 30 |
| 4.1 | Memetic Algorithms | 30 |
| 4.1.1 | Memetic Algorithms for Feature Selection | 31 |
| 4.2 | Memetic Algorithm for Multi-objective (α, β) - k -Feature Set Problem (MOMA-ABK) 32 | |
| 4.2.1 | Solution Representation | 33 |
| 4.2.2 | Initialization | 33 |
| 4.2.3 | Recombination Procedure | 34 |
| 4.2.4 | Mutation Procedure | 35 |
| 4.2.5 | Satisfaction Heuristic | 36 |
| 4.2.6 | Replacement Strategies | 39 |
| 4.2.7 | Stopping Criteria | 39 |
| 4.3 | Experiments | 40 |
| 4.3.1 | Datasets | 41 |
| 4.3.2 | Preprocessing | 42 |
| 4.4 | Computational Results | 42 |
| 4.4.1 | <i>Average</i> hypervolume | 43 |
| 4.4.2 | <i>Global Pareto Front</i> Hypervolume | 44 |
| 4.4.3 | Comparison of <i>average</i> and <i>Global Pareto front</i> Approaches. | 45 |
| 4.4.4 | Execution Time | 50 |
| 4.5 | Case of study: heat maps of biomarkers | 51 |
| 4.5.1 | Methodology | 51 |
| 4.5.2 | Results | 51 |
| 4.6 | Summary | 59 |
| 5 | The improved Version of the Memetic Algorithm for Multi-objective (α, β)-k-Feature Set Problem (MOMA-ABK) | 60 |
| 5.1 | First Improved Version of MOMA-ABK: Clustering of features | 61 |
| 5.1.1 | Feature Representation | 62 |
| 5.1.2 | Clustering-based initialization Heuristic | 63 |
| 5.1.3 | Clustering-based Satisfaction Heuristic | 63 |
| 5.2 | Experiments for Clustering-based MOMA-ABK | 65 |
| 5.2.1 | Datasets | 65 |
| 5.2.2 | Clustering Procedure | 65 |
| 5.3 | Computational Results for Clustering-based MOMA-ABK | 66 |

| | | |
|----------|---|-----------|
| 5.3.1 | Average Hypervolume | 67 |
| 5.3.2 | C-measure | 67 |
| 5.3.3 | Execution Time | 68 |
| 5.3.4 | Clustering of features | 68 |
| 5.4 | Analysis of Clustering-based MOMA-ABK results | 70 |
| 5.5 | Second improved Version of MOMA-ABK: Index of Features | 70 |
| 5.5.1 | Hierarchical <i>k-means</i> tree | 71 |
| 5.5.2 | <i>Index Initialization heuristic (IndexInit)</i> | 72 |
| 5.5.3 | <i>Index Satisfaction heuristic (IndexSH)</i> | 73 |
| 5.6 | Experiments for index-based MOMA-ABK | 73 |
| 5.6.1 | Memetic Algorithm Tuning | 74 |
| 5.6.2 | Datasets | 75 |
| 5.6.3 | <i>Index Satisfaction heuristic</i> Parameter Experiments | 75 |
| 5.6.4 | Index of Features | 76 |
| 5.7 | Computational results for index-based MOMA-ABK | 76 |
| 5.7.1 | Average hypervolume | 76 |
| 5.7.2 | Execution Time | 77 |
| 5.8 | Comparison of Clustering-based and index-based improvement | 77 |
| 5.9 | Summary | 78 |
| 6 | A Filter Feature Selection Based on the (α, β)-<i>k</i> Feature Set Problem: Microarray Datasets | 83 |
| 6.1 | Methodology for Using MOMA-ABK as a Filter Feature Selection | 83 |
| 6.1.1 | Preprocessing | 84 |
| 6.1.2 | Memetic Algorithm for Multi-objective (α, β) - <i>k</i> -Feature Set Problem (MOMA-ABK) | 85 |
| 6.1.3 | Classification | 85 |
| 6.1.4 | Postprocessing | 86 |
| 6.2 | Experiments | 87 |
| 6.2.1 | Datasets | 87 |
| 6.2.2 | Classifiers | 89 |
| 6.2.3 | Representative Pareto Front Mechanims | 89 |
| 6.3 | Computational Results for α -constraint = α^* Scenario | 89 |
| 6.3.1 | Best Memetic Algorithm | 89 |
| 6.3.2 | <i>Minimum Coverage</i> Mechanism | 91 |
| 6.3.3 | <i>Minimum Feature</i> Mechanism | 95 |
| 6.3.4 | <i>Global Pareto front</i> Mechanism | 99 |
| 6.4 | Analysis of Results from α -constraint = α^* Scenario | 103 |
| 6.4.1 | Comparison of <i>Representative Pareto front</i> Mechanisms | 103 |
| 6.4.2 | Comparison with Filter Feature Selection State-of-the-art Algorithms | 107 |
| 6.5 | A less restricted scenario: α -constraint=1 | 108 |
| 6.5.1 | <i>Minimum Coverage</i> Mechanism | 108 |
| 6.5.2 | <i>Minimum Feature</i> Mechanism | 113 |

| | | |
|----------|---|------------|
| 6.5.3 | <i>Global Pareto front</i> mechanism | 117 |
| 6.6 | Analysis of Results from α - constraint =1 Scenario | 121 |
| 6.6.1 | Comparison of <i>Representative Pareto front</i> mechanisms | 121 |
| 6.6.2 | Comparison with Filter Feature Selection State-of-the-art algorithms | 122 |
| 6.6.3 | Comparison with Unfiltered Dataset | 126 |
| 6.6.4 | Comparison of the Performance of Different Classifiers | 127 |
| 6.7 | Summary | 128 |
| 7 | Conclusions and Future work | 130 |
| 7.1 | Summary of Contributions | 131 |
| 7.2 | Future research | 132 |
| | References | 134 |

List of Tables

| | | |
|-----|--|----|
| 2.1 | Input data example composed of four samples with five features and the class value. Matrix \mathcal{D} of 4×5 and the \mathcal{T} of 4×1 | 10 |
| 2.2 | Complete set of possible subset of features for a dataset with five features. | 10 |
| 2.3 | Complete set of possible solutions for five features. | 13 |
| 2.4 | Coverage for each feature and each node in the graph representation of the dataset example. The nodes have only one coverage and it is the number of features connected. In contrast, the features have an α coverage, which is the number of α Nodes connected; an β coverage, which is the number of β Nodes connected; and <i>Coverage</i> to indicate the sum of the previous coverages. | 15 |
| 4.1 | Parameters of the algorithm. | 40 |
| 4.2 | Description of datasets used, showing the number of features, samples, α Nodes, and the ratio between features and samples. | 41 |
| 4.3 | Hypervolume Normalized for each algorithm (MA) considering 6 datasets. The value presented is the average hypervolume normalized across 30 trials. | 43 |
| 4.4 | Wilcoxon statistical test between satisfaction heuristics. In this table, we present the <i>p-value</i> between the two satisfaction heuristics and the algorithm without any fixing heuristic when we test the differences across the six datasets. | 44 |
| 4.5 | Hypervolume normalized of the global Pareto front. For each algorithm, we present the hypervolume normalized of the global Pareto front obtained from 30 trials. | 45 |
| 4.6 | Wilcoxon statistical test between satisfaction heuristics. This table presents the <i>p-value</i> between the two satisfaction heuristics and the algorithm without any fixing heuristic when we test the differences across the six datasets. | 45 |
| 4.7 | Average execution time for each algorithm configuration. For each dataset, we present the average execution time across the 30 trials. | 50 |
| 4.8 | Wilcoxon statistical test between satisfaction heuristics. This table, we present the <i>p-value</i> between the two satisfaction heuristics and the algorithm without any fixing heuristic when we test the differences across the six datasets. | 50 |

| | | |
|------|--|-----|
| 5.1 | Feature representation for clustering procedure. The representation of each feature is an expression of $\alpha Nodes$, where 1 means the feature is connected in the graph with the node, and the value 0 means otherwise. | 62 |
| 5.2 | Parameters of the memetic algorithms. The referenced algorithms are detailed at Section 4.2. | 65 |
| 5.3 | Average hypervolume for each algorithm across 30 trials for each dataset. | 67 |
| 5.4 | Average c-measure between the four algorithms. | 68 |
| 5.5 | Average processing time in seconds. Time in seconds is presented by the average value across 30 trials for each dataset. | 68 |
| 5.6 | Details about the clustering of features. | 68 |
| 5.7 | Parameters of the algorithm. | 75 |
| 5.8 | Number of features used in the <i>IndexSH</i> . For each dataset and experiment, we present the number of features fetched in the satisfaction heuristic based on the index. | 75 |
| 5.9 | Average hypervolume across six neighborhood sizes for six datasets. For each dataset and neighborhood size we present the average hypervolume across 30 trials. | 79 |
| 5.10 | Best configuration algorithm for every neighborhood size in the implemented Index satisfaction heuristic. For each database, we present the initialization procedure, the recombination and mutation operators of the algorithm with the highest average hypervolume. | 80 |
| 5.11 | Average execution time across six neighborhood sizes for six datasets. For each dataset and neighborhood size we present the average execution time across 30 trials | 81 |
| 5.12 | Average hypervolume for each algorithm across 30 trials for each dataset (<i>Intersect</i> recombination and <i>DetBitFlip(3)</i> mutation). | 82 |
| 6.1 | Parameters of the algorithm. | 87 |
| 6.2 | Summary characteristics for the datasets. | 87 |
| 6.3 | Summary characteristics for the datasets after Fayyad and Irani heuristic. | 89 |
| 6.4 | Average normalized hypervolume using 30 trials. For each dataset, we present the hypervolume in the 36 algorithms. For each dataset, the highest hypervolume is highlighted. | 90 |
| 6.5 | Best algorithm configuration for each dataset. For each best configuration, we present the initialization, the mutation, the recombination and the satisfaction heuristic used during the experiments. In addition, we present the average of the normalized hypervolume for each dataset. | 91 |
| 6.6 | Current single-objective state-of-the-art algorithms. In this table, we present per each dataset the accuracy and the number of features (between parenthesis) reported in the study respectively. | 122 |

| | | |
|-----|--|-----|
| 6.7 | Classification accuracy without filter feature selection. For each dataset, we present the average classification accuracy considering the four classifiers. In addition, we present the average number of features in each dataset after the application of the Fayyad and Irani heuristic. | 126 |
| 6.8 | Classification accuracy of using all the features for each dataset. For each dataset, the accuracy of each classifier is presented for each of the three training-test pair of files. | 127 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Filter method basic flowchart. This figure presents the two stages of a filter feature selection methodology. The algorithm first determines a subset of features able to perform the separation of the class values, and then evaluates empirically the classification accuracy of the subset of features. | 8 |
| 2.2 | Wrapper method basic flowchart. The figure presents the two stages of a wrapper feature selection method. The algorithm determines the subset of features that has the highest classification accuracy over a sample of the original dataset. The algorithm then uses the selected subset of features to determine the classification accuracy using an unseen sample of the dataset. | 9 |
| 2.3 | Graph representation of the input dataset of Table 2.1. For each feature exists a node (F_1, F_2, F_3, F_4 and F_5), $\alpha Nodes$ are : $\{(1, 2); (2, 3); (2, 4)\}$ and $\beta Nodes$ are: $\{(1, 3); (1, 4); (3, 4)\}$. The feature has an edge with an $\alpha Node$ if the feature has different values in the samples of the $\alpha Node$. Complementary, the feature has an edge with a $\beta Node$ if the feature has the same value in the samples of the $\beta Node$ | 12 |
| 3.1 | Solution Space and Pareto front for multi-objective the problem regarding purchasing a house. In the figure, the solution space is composed by all the dots in the space. The black dots represent units, pink dots represent town houses, green dots represent houses and dark pink dots represent mansions. The Pareto front is represented by the solutions associated with pictures. | 20 |
| 3.2 | Pareto dominance between solutions in a minimization-maximization problem. The blue square is weakly dominated by the red square (same value of Objective 1) and by the green square (same value in Objective 2). There is non-dominance between the green and red squares. | 22 |
| 3.3 | Pareto front for a (minimization, maximization) multi-objective optimization problem. Point A is the worst point based on the Pareto front, and has the worst known value for all objectives. | 24 |
| 3.4 | Hypervolume of a Pareto front for a (minimization, maximization) multi-objective optimization problem. The red area represents the hypervolume generated by the <i>paretoA</i> and the blue area represents the hypervolume generated by the <i>paretoB</i> | 25 |

| | | |
|-----|--|----|
| 4.1 | Representation of a solution inside our algorithm. The j -bit represents the j -feature, where the value of 1 represents that the feature belongs to the subset of features S (i.e., $S[j] = 1$), and the value of 0 represents that the feature is not in the solution (i.e., $S[j] = 0$) | 33 |
| 4.2 | Example of a solution considering our solution representation. The solution composed by the subset of features $\{F1, F4, F5\}$ has the bit-array representation of $S[1] = 1, S[2] = 0, S[3] = 0, S[4] = 1, S[5] = 1$ | 33 |
| 4.3 | Example of values utilized during greedy satisfaction heuristic. The features with bold borders represent the features inside the solution, and the features with thin borders are the <i>NotIncluded</i> set. The dark green α Nodes are unsatisfied. The bold border α Nodes are the <i>minimum-αNodes</i> . There are blue edges between the possible features and <i>minimum-αNodes</i> | 38 |
| 4.4 | Example of values used during cover satisfaction heuristic. The features with bold borders represent the features inside the solution, and the features with thin borders are the <i>NotIncluded</i> set. The dark green α Nodes are unsatisfied. There are blue edges between the possible features and unsatisfied α Nodes. | 40 |
| 4.5 | Global Pareto front and five Pareto front for an algorithm continuation. The global Pareto front is represented by the black line and the solution Pareto front are colored. The x-axis is the number of features in the subset (k), and the y-axis is the β value of the solution. | 47 |
| 4.6 | Global Pareto front and five Pareto front for an algorithm continuation. The global Pareto front is represented by the black line and the solution Pareto front are colored. The x-axis is the number of features in the subset (k), and the y-axis is the β value of the solution. | 48 |
| 4.7 | Global Pareto front and five Pareto front for an algorithm continuation. The global Pareto front is represented by the black line and the solution Pareto front are colored. The x-axis is the number of features in the subset (k), and the y-axis is the β value of the solution. | 49 |
| 4.8 | Heatmap for DownSyn dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 53 |

| | | |
|------|---|----|
| 4.9 | Heatmap for Smoking dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 54 |
| 4.10 | Heatmap for Bruta dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 55 |
| 4.11 | Heatmap for PdParkinson dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 56 |
| 4.12 | Heatmap for Prostate dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 57 |
| 4.13 | Heatmap for Parkinson dataset. On the left side the heatmap of the solution from the Pareto front with the maximum number of features, on the right side the heatmap of the solution with the minimum number of features found in the Pareto front. The x-axis are the genes (features), and the y-axis are the subjects (samples). The label of the sample is blue if the sample does not have the disease (class value = 0), and the label is black if the sample has the disease (class value = 1). The global Pareto front is represented by the black line and the solution Pareto front are colored. | 58 |

| | | |
|-----|--|----|
| 5.1 | Representation of a clustering technique. Each color represent a different pattern structure in the data points. The group <i>centroids</i> are represented by the points with double black contoured line. | 61 |
| 5.2 | A graph representation of an instance of the ABkFS problem only considering the $\alpha Nodes$ side. Features are $(f_1, f_2, f_3, f_4$ and $f_5)$ and $\alpha Nodes$ are $\{(1, 2); (2, 3); (2, 4)\}$ | 62 |
| 5.3 | Flowchart of the clustering process. The process first needs to determine the number of clusters. Therefore, we create a dendogram of the dataset, and we determine the level where the variance of the cluster is the minimum. In the second part, we use the number of clusters of the level selected in the previous stage, and we perform <i>k-means</i> algorithm to define the clusters to be used in our algorithms. | 66 |
| 5.4 | Frequency distribution of clusters size. For each dataset, the figure represents the number of clusters with a specific cluster size. The x-axis is the frequency or number of clusters, and the y-axis is the size of the cluster represented by the number of features. The blue line represents the Bruta dataset, the black line represents the Parkinson dataset, the purple line represents the Downsyn dataset, the green line represents the Prostate dataset, the blue line represents the Smoking dataset and the red line represents the PdParkinson dataset. | 69 |
| 5.5 | Hierarchical k-means index. Each color separation represents a level in the hierarchy. The first level is pink and generates four groups ($b_R = 4$). To obtain the second level, we apply <i>k-means</i> for each group and obtain the purple separation. | 71 |
| 6.1 | Flowchart for classification task using MOMA-ABK as a Filter Feature Selection. The proposed filter feture selection has four stages: preprocessing, MOMA-ABK , classification, and postprocessing. The interaction between the stages are represented by dotted lines, and each colored line represents a file from the three-floed cross validation. In the last stage postprocessing, the figure presents the three different alternatives to generate the <i>Representative Pareto front</i> | 84 |
| 6.2 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 92 |

| | | |
|-----|---|-----|
| 6.3 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 93 |
| 6.4 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 94 |
| 6.5 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 96 |
| 6.6 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 97 |
| 6.7 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 98 |
| 6.8 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 100 |
| 6.9 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 101 |

| | | |
|------|---|-----|
| 6.10 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 102 |
| 6.11 | Comparison between the three <i>Representative Pareto front</i> mechanisms. For each dataset, we present the three average hypervolume generated by each mechanism. In order to have a representative behaviour, we average the hypervolume obtained considering the four different classifiers. The blue bar represents the MC mechanism, the red bar represents the MF mechanism, and the yellow bar represents the GPF mechanism | 103 |
| 6.12 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 104 |
| 6.13 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 105 |
| 6.14 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 106 |
| 6.15 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 110 |
| 6.16 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 111 |
| 6.17 | Minimum Coverage mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 112 |

| | | |
|------|---|-----|
| 6.18 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 114 |
| 6.19 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 115 |
| 6.20 | Minimum Feature mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 116 |
| 6.21 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 118 |
| 6.22 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 119 |
| 6.23 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers. The pink front represents the J48 classifier, the red front represents KNN classifier, the green front represents the SVM classifier and the blue front represents the NB classifier. The x-axis represents the number of features and the y-axis is the accuracy of the classifier using the subset of features. | 120 |
| 6.24 | Comparison between the three <i>Representative Pareto front</i> mechanisms. For each dataset, we present the three average hypervolume generated by each mechanism. In order to have a representative behaviour, we average the hypervolume obtained considering the four different classifiers. The blue bar represents the MC mechanism, the red bar represents the MF mechanism, and the yellow bar represents the GPF mechanism | 121 |

| | | |
|------|--|-----|
| 6.25 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 123 |
| 6.26 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 124 |
| 6.27 | Global Pareto Front mechanism. For each dataset, we present the Representative Pareto front considering the four different classifiers (represented by lines). In addition, we present the state-of-the-art algorithms, as points in the figure, indicating the number of features and the accuracy reported. . | 125 |
| 6.28 | Hypervolume of Representative Pareto front from GPF mechanism. For each dataset and classifier, we present the hypervolume of the Representative Pareto front presented in Figure 6.25. The pink bar corresponds to J48 classifier, the red bar represents KNN, the blue bar represents the SVM classifier and the NB is represented by the green bar. | 128 |

List of Algorithms

| | | |
|----|---|----|
| 1 | Pseudocode to calculate α^* value. | 28 |
| 2 | Pseudocode of the Memetic Algorithm. | 31 |
| 3 | RandomInit pseudocode, random initialization algorithm. | 34 |
| 4 | Intersect recombination pseudocode. | 35 |
| 5 | Uniform crossover(UXovr) pseudocode. | 35 |
| 6 | DetBitFlip mutation pseudocode. | 36 |
| 7 | Bit mutation pseudocode. | 36 |
| 8 | <i>BitInvert</i> mutation pseudocode. | 37 |
| 9 | Pseudocode of <i>Greedy Satisfaction heuristic (GreedySH)</i> | 37 |
| 10 | Pseudocode of <i>Cover Satisfaction heuristic (CoverSH)</i> | 39 |
| 11 | <i>Cluster Initialization heuristic</i> , clustering-based initialization pseudocode. | 63 |
| 12 | <i>Cluster Satisfaction heuristic</i> , clustering-based satisfaction heuristic pseudocode. | 64 |
| 13 | Cluster assignation pseudocode. | 72 |
| 14 | <i>Index Initialization heuristic</i> , index-based initialization pseudocode. | 73 |
| 15 | <i>Index Satisfaction heuristic</i> pseudocode, index-based satisfaction heuristic. | 74 |

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
S00 single-objective optimization. 26
SOP single-objective optimization problem. 19, 21
S0 Single-objective. 4, 24
S0 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*