Efficient Methods of Feature Selection Based on Combinatorial Optimization Motivated by the Analysis of Large Biological Datasets

Mateus Rocha de Paula
M.Sci. (Computer Science)
B.Sci. (Computer Science)

This dissertation is submitted as a partial requirement for the Degree of Doctor of Philosophy

Faculty of Engineering and Built Environment
School of Electrical Engineering and Computer Science

Newcastle, NSW, Australia
August, 2012
Statement of Originality

The thesis contains no material which has been accepted 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 in the text. I give consent to this copy of my thesis, when deposited in the University Library\(^1\), being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

Mateus Rocha de Paula

\(^1\)Unless an Embargo has been approved for a determined period.
Matsuo Basho, one of the most celebrated Japanese poets, once wrote “Every day is a journey, and the journey itself is home”. When I think about my PhD today, that couldn’t sound more true. The degree of Doctor of Philosophy not only represents the completion of this work, but also celebrates a series of meaningful events that allowed me to grow both as a person and professional; and marks the beginning of a new and exciting stage of my life. Every little step I took since I ventured this path has its own story, of immense importance, and none of which would be possible without the help great personalities that I was already fortunate to know, and others that I had the great pleasure to meet on my way. To all those, my many big thanks.

To my mum France, dad José Vicente and sister Rafaella, my eternal gratitude for all the love and support. Every time the light in the end of the tunnel seemed a little distant, it was you who made it brighter. This victory is very much yours too.

Many thanks to my supervisors Pablo and Regina, who so patiently mentored me through all the process, teaching me the skills and giving me the tools that I needed to succeed on this challenge and many yet to come. You not only helped me to be a better researcher, but also taught me how to be truly committed to science, and for that I’ll always be thankful. To my friend and colleague Martín for supporting me from the very beginning of my academic career and for opening the doors to this opportunity, to Australia and of his own home to me, making it all possible. It’s always been a pleasure to work with such a good friend, always with an open mind, positive attitude and a smile on his face! A big thanks to my friend Alexandre, for all the help both inside and outside the lab. This story would have had a much different ending without your advice, friendship, and many beers and fun times shared. To
my friend and colleague Lee-Anne, for showing me the ways of her people, helping me to feel at home on either continent! To my friend and colleague Carlos also a big thanks, for all the help, tips, teachings and laughter. It is really inspiring to see such a curious and enthusiastic mind working! Even brewing and drinking beer is science around you, and it is good science! And to my friends and colleagues Renato and Ahmed, who shared the path with me, with its challenges and joys.

I’d also like to thank the panel of examiners that reviewed this document. All your remarks were very helpful and constructive. They certainly helped me to improve my thesis, and for that - and also the acceptance and compliments, of course - I thank you!

Last but not least, I’d also like to dedicate this thesis to all my loyal friends, from all over the world. To my country fellows, for always being there for me, reminding me that I’ll always have a safe place by their side either at home or from a distance. And to all my new ones as well, for welcoming me to their lands and hearts, giving me the strength to be a citizen of the world.
To Vó Cota, Tio Tonico and Shilly.

Wherever you are.
Contents

1 Introduction 1

1.1 Bioinformatics and Learning Algorithms 2
1.2 Bioinformatics and Optimization 3
1.3 Feature Selection 6
1.4 Research Questions 8
1.5 Structure of This Thesis 8

2 Existing Feature Selection Approaches 11

2.1 Wrapper Methods 12
2.2 Embedded Methods 13
2.3 Filter Methods 13

2.3.1 Univariate Methods 13
2.3.2 Multivariate Methods 14

2.4 Further Reading 15

3 The \((\alpha, \beta)\)-k-Feature Set Problem Approach 17

3.1 The Problem Formulation 18

3.1.1 The Graph Representation 18

3.2 The Feature Selection Approach 19

3.3 Set (Multi) Cover Problems 28

3.3.1 Existing Set Cover Approaches 31
3.3.2 Existing Set Multi Cover Approaches 33
5.6.3 Data Structures ......................................................... 90
5.6.4 Caching ................................................................. 90
5.6.5 Parallelization ......................................................... 91
5.7 Conclusions ............................................................... 92

6 Solution Quality and Dual Bounds ........................................ 93
   6.1 Greedy Dual Bounds ................................................ 95
      6.1.1 Lower bound for Min k ........................................ 95
      6.1.2 Upper bound for Max $\beta$ ................................ 96
      6.1.3 Upper bound for the Max Total Covering .................. 96
   6.2 Lagrangian Relaxation .............................................. 97
      6.2.1 The Subgradient Method ..................................... 98
      6.2.2 Lower Bounds for Min k ..................................... 99
      6.2.3 Upper Bounds for Max $\beta$ ................................. 100
      6.2.4 Upper Bounds for Max Cover ................................ 102
      6.2.5 Implementation Remarks .................................... 102
      6.2.6 Lagrangian Heuristics and Variable Fixing ................ 104
   6.3 Other Alternatives .................................................. 105
   6.4 Conclusions .......................................................... 108

7 Computational Results .................................................... 109
   7.1 Primal Heuristics .................................................... 109
      7.1.1 Greedy Algorithms .......................................... 110
      7.1.2 Local Searches .............................................. 110
      7.1.3 Meta-Heuristics ............................................. 112
   7.2 Dual Bounds .......................................................... 125
      7.2.1 Lower bounds of Min $k$ .................................... 125
      7.2.2 Upper bounds of Max $\beta$ ................................. 127
      7.2.3 Upper bounds of the Max Total Covering .................. 127
      7.2.4 Variable Fixing ............................................. 128
8 Conclusions

8.1 Primal Heuristics ........................................ 129
8.2 Dual Bounds ............................................. 132
8.3 Concluding Remarks and Future Research ............. 134

A Summary of Notation ....................................... 137

B Discretization and Pre-Selection of Features ............ 139

C Case-Control Datasets: Valid and Invalid Graph Extracts 141

D Result Listings .............................................. 147

D.1 Chapter 4 ...................................................... 147
D.2 Chapter 5 ...................................................... 153
D.3 Chapter 6 ...................................................... 161
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>An $(\alpha, \beta)$-k-Feature Set Problem instance example and its graph representation.</td>
<td>20</td>
</tr>
<tr>
<td>3.2</td>
<td>Mathematical formulation of the $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>The four-stage approach used to determine the $(\alpha, \beta)$-k-Feature Set Problem parameters, $\alpha^<em>, \beta^</em>$ and $k^*$.</td>
<td>23</td>
</tr>
<tr>
<td>3.4</td>
<td>Mathematical formulation of the Max $\alpha$ $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>24</td>
</tr>
<tr>
<td>3.5</td>
<td>Mathematical formulation of the Min $k$ $(\alpha, \beta)$-k-Feature Set Problem with $\beta = 0$.</td>
<td>25</td>
</tr>
<tr>
<td>3.6</td>
<td>Mathematical formulation of the Max $\beta$ $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>26</td>
</tr>
<tr>
<td>3.7</td>
<td>Mathematical formulation of the Max Cover $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>27</td>
</tr>
<tr>
<td>3.8</td>
<td>Example of a Min Set Cover.</td>
<td>29</td>
</tr>
<tr>
<td>4.2</td>
<td>Probability Density Function of the Linear Distribution.</td>
<td>51</td>
</tr>
<tr>
<td>4.3</td>
<td>Running times of exact approach.</td>
<td>56</td>
</tr>
<tr>
<td>4.4</td>
<td>Exact approach gaps: Min $k$.</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Exact approach gaps: Max $\beta$.</td>
<td>59</td>
</tr>
<tr>
<td>4.6</td>
<td>Exact approach gaps: Max Cover.</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Movement example.</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Affected sample pairs after swapping 2 features.</td>
<td>88</td>
</tr>
<tr>
<td>5.3</td>
<td>Affected sample pairs after swapping 2 features for 1.</td>
<td>89</td>
</tr>
<tr>
<td>6.1</td>
<td>Lagrangian Relaxation of the Min $k$ $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>99</td>
</tr>
<tr>
<td>6.2</td>
<td>Lagrangian Relaxation of the Max $\beta$ $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>101</td>
</tr>
<tr>
<td>6.3</td>
<td>Lagrangian Relaxation of the Max Cover $(\alpha, \beta)$-k-Feature Set Problem.</td>
<td>103</td>
</tr>
</tbody>
</table>
6.4 Mathematical formulation of the dual problem of the Max $\alpha$ $(\alpha, \beta)$-k-Feature Set Problem. .................................................. 106
6.5 Mathematical formulation of the dual problem of the Min $k$ $(\alpha, \beta)$-k-Feature Set Problem. .................................................. 106
6.6 Mathematical formulation of the dual problem of the Max $\beta$ $(\alpha, \beta)$-k-Feature Set Problem. .................................................. 107
6.7 Mathematical formulation of the dual problem of the Max cover $(\alpha, \beta)$-k-Feature Set Problem. .................................................. 107

7.1 Optimality gaps: MH1 ........................................................... 116
7.2 Optimality gaps: MH2 ........................................................... 117
7.3 Optimality gaps: MH3 ........................................................... 118
7.4 Optimality gaps: MH4 ........................................................... 119
7.5 Optimality gaps: MH5 ........................................................... 120

C.1 Valid (realistic) graph patterns for the $(\alpha, \beta) - k$ Feature Set problem, compared to the Set (Multi) Cover problem. .................................................. 143
C.2 Invalid (unrealistic) graph patterns for the $(\alpha, \beta) - k$ Feature Set problem, compared to the Set (multi)Cover problem. .................................................. 145
List of Tables

4.1 Exact results for real world instances ........................................ 40
4.2 Proposed instances generation parameters ................................... 54
4.3 Average CPLEX Results .............................................................. 58

7.1 Summary of Algorithms ............................................................. 111
7.2 Combined Algorithms ............................................................... 112
7.3 Greedy Heuristic Comparison: results for each instance class. ............ 113
7.4 Local Search Impact Comparison: Results of Redundancy Tests (First Improvement) ................................................................. 114
7.5 Local Search Impact Comparison: Results of Redundancy Tests (Best Improvement) ................................................................. 115
7.6 Average primal Gaps and Running times ......................................... 122
7.7 Individual Stage Comparison. ......................................................... 122
7.8 Results of the heuristics for real world instances .................................. 123
7.9 Running times of the heuristics for real world instances .................... 124
7.10 Dual Bounds for Min k ................................................................. 126
7.11 Average Dual Gaps and Running times for the maximisation of $\beta$ ........ 127
7.12 Average Dual Gaps and Running times for the maximisation of the Total Covering 128

D.1 CPLEX times ............................................................................. 148
D.3 CPLEX results Max $\beta$ .............................................................. 150
D.2 CPLEX results Min k ................................................................. 151
D.4 CPLEX results Max Cover .............................................. 152
D.5 Metaheuristics’ Results .................................................. 154
D.6 Metaheuristics’ Times ...................................................... 159
D.7 Dual Bounds for Max $\beta$ ............................................... 162
D.8 Dual Bounds for Max Cover ............................................. 165
List of Algorithms

1. Polynomial algorithm to solve the Max $\alpha (\alpha, \beta)$-k-Feature Set Problem. 23
2. $(\alpha, \beta)$-k-Feature Set Problem Instance Reduction Rule 1 35
3. $(\alpha, \beta)$-k-Feature Set Problem Instance Reduction Rule 2 35
4. $(\alpha, \beta)$-k-Feature Set Problem Instance Reduction Rule 3 35
5. $(\alpha, \beta)$-k-Feature Set Problem Instance Reduction Rule 4 36
6. Instance Generator 49
7. Linear Random Number Generator 50
8. Continuous to Integer 50
9. Instance Generator: fix disconnections 52
10. Greedy Algorithm 1 69
11. Greedy Algorithm 2 69
12. Greedy Algorithm 3 70
13. A swap-based first-improvement local search 72
14. A swap-based best-improvement local search 73
15. Randomization method focused on the greedy algorithm 75
16. Randomization method focused on the local search 75
17. A randomized swap-based first-improvement local search 77
18. Typical VNS algorithm 78
19. Modified VNS algorithm 81
20. Modified VNS algorithm With Tabu Search 84
21. Randomization method focused on the greedy algorithm, considering Tabu Search 85
22. Randomization method focused on the local search, considering Tabu Search 85
<table>
<thead>
<tr>
<th>No.</th>
<th>Algorithm</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Greedy Algorithm 1 with a tabu list</td>
<td>86</td>
</tr>
<tr>
<td>24</td>
<td>Greedy Algorithm 2 with a tabu list</td>
<td>86</td>
</tr>
<tr>
<td>25</td>
<td>Greedy Algorithm 3 with a tabu list</td>
<td>87</td>
</tr>
<tr>
<td>26</td>
<td>Subgradient algorithm</td>
<td>98</td>
</tr>
</tbody>
</table>
Abstract

Intuitively, the Feature Selection problem is to choose a subset of a given set of features that best represents the whole in a particular aspect, preserving the original semantics of the variables on the given samples and classes. In practice, the objective of finding such a subset is often to reveal a particular characteristic present in the given samples.

In 2004, a new feature selection approach was proposed. It was based on a combinatorial optimization problem called $(\alpha, \beta)$-k-Feature Set Problem. The main advantage of using this approach over ranking methods is that the features are evaluated as groups, instead of only considering their individual performance.

The main drawback of this approach is the complexity of the combinatorial problems involved. Since some of them are NP-Complete, it is unlikely that there would exist an efficient method to solve them to optimality efficiently. To the best of the author’s knowledge at the moment of this research, the available tools to deal with the $(\alpha, \beta)$-k-Feature Set Problem approach can not solve problems of the magnitude required by many practical applications.

Given the big advantage brought by the multivariate characteristic of this method, its successful wide applicability and knowing that its only real known drawback is scalability, further research to overcome such a difficulty is appropriate. Even though the optimal solution of the problem is always desirable, it often is not strictly necessary in the case of many biological applications. Therefore, this work aims to propose fast heuristics to address the $(\alpha, \beta)$-k-Feature Set Problem approach, and propose procedures to obtain dual bounds that do not rely on external optimization packages.