

Mining disjunctive patterns in biomedical data sets

Renato Vimieiro

M.Sci. (Computer Science)

B.Sci. (Computer Science)

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



THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

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

Newcastle, NSW, Australia
August, 2012

Statement of Originality

This 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, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

Statement of Authorship and Collaboration

I hereby certify that the work embodied in this thesis contains a published paper of which I am a joint author. Chapter 3 contains parts of the paper “Mining disjunctive minimal generators with TitanicOR” published on the journal *Expert Systems with Applications*, Volume 39, Issue 9 (Vimieiro and Moscato, 2012). This is the result of a collaborative work with Prof. Pablo Moscato, my principal supervisor, in an ordinary student-supervisor collaboration practice. Nevertheless, it is worth mentioning that I had an active role in every stage of the development of such work. I included this written statement, endorsed by my supervisor, attesting to my contribution to the aforesaid paper.

Renato Vimieiro

Prof. Pablo Moscato

*to Ana, my wife, for her love, patience,
and support along this intense journey*

Acknowledgements

Despite having dedicated this thesis to my wife Ana, my first acknowledgement must go to her as well. I want to thank her for all the support that she gave me during these long forty months of my research. She not only gave me strength and motivation during the moments I felt desolated with this thesis, but our talks about how research should be conducted also helped me improve the quality of my work. Thanks Ana. Hopefully, in three years it will be your moment!

Secondly, I would like to thank my supervisors, Pablo and Regina, for everything they did for me during my stay in CIBM. Pablo more than anyone bet on me, a person who lived in the other side of the world, who he didn't know, and, in spite of that, helped covering part of my scholarship. I hope I corresponded to your expectations, Pablo.

Apart from financial support, which, nevertheless, is extremely important, Pablo and Regina both taught me loads of things that I will carry for the rest of my life. Pablo is a very enthusiastic researcher, one of the best persons I have seen for brainstorming and producing insightful ideas. Go talk to him for five minutes and he will give you heaps of excellent ideas. This ability is very useful for PhD candidates like me, who sometimes may run out of ideas and get stuck. Regina, by the other hand, was very helpful to keep me on track. She is a very organized person and, although we didn't interact as much as I did with Pablo (since she was my co-supervisor), she managed to put Pablo and I on track, stop brainstorming, start executing my ideas, and finish with this PhD. Therefore, she was also fundamental for my learning. I believe discipline and creativity are very important for PhDs — of course, that's for anyone, but this is a PhD Dissertation; so, excuse me from putting it that way — and thankfully I could find this in these two supervisors. Thanks very much Pablo and Regina!

Thirdly, I must thank CIBM's Argentinian wing for everything they made for me. Starting with Martin Ravetti who introduced me to Pablo and gave me motivation to apply for the PhD position. Then, Osvaldo Rosso, an amazing person I had the pleasure of meeting here in Newcastle. Osvaldo hosted me in his house for two months when I first arrived in Australia. He also helped me to set up my (at that time) new home here and he did all of that without even knowing me. I am extremely thankful for that, Osvaldo! Finally, Carlos Riveros. I have to thank Carlos for so many things that I think I'll have to write a dissertation only with that! Not only he gave me several bottles of his home-brew — which I confess, didn't last as long as I expected —, he also helped me with some nasty programming/research problems. He helped me, for instance, advance my knowledge on CUDA/GPU programming, which I had never heard before coming to Australia. He also gave me good advices on what I should do with/for my work. Seriously, I have no words to thank Carlos for the help he gave me during my PhD.

Then, I have to thank all my friends, colleagues, and especially my family for their friendship and support during these years. Thanks my Australo-Brazilian friends Tim and Andreza Ireland, Mateus Rocha de Paula, Alexandre Mendes, Ahmed Shamsul Arefin (who opened my eyes for loads of interesting things about the Islamic and Bengali cultures), Lee-Anne Marsh (who was very helpful to me in my first year and showed me how hospitable Australians can be), my parents Edna and Laurindo, and also to my mother-in-law Rita. Thanks to all of you guys.

Finally, I am very thankful to the University of Newcastle for my scholarships and for the opportunity to be part of an incredible academic environment. My Doctorate was fully funded by one University of Newcastle Postgraduate Research Scholarship and a University of Newcastle International Postgraduate Research Scholarship. Thanks very much.

To all the others that are not covered on this lengthy statement, please forgive my rudeness and be sure you were also important in my journey.

Contents

1	Introduction	1
1.1	Our research approach and objectives	5
1.2	Thesis organization	6
2	Literature Review	7
2.1	Data Mining and the KDD	7
2.2	Frequent pattern mining	9
2.3	Concise representations for frequent patterns	11
2.3.1	Maximal frequent itemsets	12
2.3.2	Frequent closed itemsets and minimal generators	12
2.3.3	Other types of frequent patterns	17
2.4	Summary	18
3	Mining disjunctive minimal generators	21
3.1	Introduction	21
3.2	Foundations	23
3.3	TitanicOR	28
3.4	Experimental results	33
3.5	Related work	39
3.5.1	Other methods to mine general association rules	39
3.5.2	Other methods to mine disjunctive patterns	42
3.5.3	Methods and applications in other domains	43
3.6	Conclusion and Future Works	44
4	Mining Disjunctive Closed Itemset	47
4.1	Introduction	47
4.2	Basic concepts	49
4.2.1	Disjunctive closed itemsets	49
4.2.2	Finding frequent disjunctive closed itemsets from samples	53
4.2.3	Conditional tables	55
4.2.4	Further reducing the search space	58
4.3	Disclosed	59
4.4	Practical issues	64
4.4.1	How we implemented Disclosed	64
4.4.2	Experiments: assessing the performance	65
4.5	Related work	75
4.5.1	Vertical itemset mining	75
4.5.2	Pattern mining in bioinformatics	77
4.5.3	Relations with conjunctive itemset mining	78

4.6	Conclusions	81
5	Mining Quasi-CNF Emerging Patterns	85
5.1	Introduction	85
5.2	Quasi-CNF emerging patterns	87
5.3	Hypergraphs, minimal transversals, and QCEPs	91
5.4	Algorithms for enumerating minimal transversals	95
5.5	Assessing the performance of the methods	96
5.5.1	Performance with good data sets	102
5.5.2	Performance with average data sets	105
5.5.3	Performance with bad data sets	106
5.5.4	Size of borders	108
5.6	Related work	109
5.6.1	Classification via association rules	109
5.6.2	Classification via emerging patterns	113
5.7	Conclusions	115
6	Conclusions and final remarks	119
6.1	Limitations	122
6.2	Open problems and future works	124
	Appendix A	141
A.1	Data sets used to assess Disclosed’s performance	141
A.2	Complementary performance figures for Disclosed	153
A.2.1	Good data sets	153
A.2.2	Bad data sets	156
A.2.3	Average data sets	159
A.3	Disclosed’s performance results	161
A.3.1	Good data sets	161
A.3.2	Bad data sets	168
A.3.3	Average data sets	172
	Appendix B	175
B.1	Tables with the performance results of QCEP	175
B.1.1	Good data sets	175
B.1.2	Average data sets	181
B.1.3	Bad data sets	187
B.2	Complementary performance figures for QCEP	191
B.2.1	Good data sets	191
B.2.2	Average data sets	194
B.2.3	Bad data sets	196

List of Figures

2.1	Maximal itemsets from the data set in Table 2.1	13
2.2	Closed itemsets from the data set in Table 2.1	14
3.1	Performance of the algorithms varying number of samples, features, and density	35
3.2	Number of itemsets mined by the algorithms with different densities .	36
3.3	Performance of the algorithms in Test 1 with distinct data sets	38
3.4	Performance of the algorithms in Test 2 with distinct data sets	40
3.5	Performance of the algorithms in Test 3 with distinct data sets	41
4.1	Enumeration of sets of samples from Table 4.1a	57
4.2	Performance measures for GDS2821, a representative of the class of data sets with which Disclosed has good performance	70
4.3	Performance measures for Lymphoma, a representative of the class of data sets with which Disclosed has bad performance	72
4.4	Performance measures for ALL-AML, a representative of the class of data sets with which Disclosed has average performance	74
5.1	Graphical example of a hypergraph built from the data set on Table 5.1	92
5.2	Performance of the algorithms when mining QCEPs in <i>good</i> data sets	104
5.3	Performance of the algorithms when mining QCEPs in <i>average</i> data sets	105
5.4	Performance of the algorithms when mining QCEPs in <i>bad</i> data sets .	107
A.1	Performance measures for Disclosed with the data set GDS963	153
A.2	Performance measures for Disclosed with the data set GDS2200	154
A.3	Performance measures for Disclosed with the data set GDS2545	154
A.4	Performance measures for Disclosed with the data set GDS2941	155
A.5	Performance measures for Disclosed with the data set Embryo	156
A.6	Performance measures for Disclosed with the data set Promoters	157
A.7	Performance measures for Disclosed with the data set GDS2519	158
A.8	Performance measures for Disclosed with the data set GDS2250	159
A.9	Performance measures for Disclosed with the data set Colon	160
A.10	Performance measures for Disclosed with the data set Leukemia	160
B.11	Performance of the algorithms with the data set ALL-AML discretized with equal frequency	191
B.12	Performance of the algorithms with the data set Colon discretized with equal frequency	192
B.13	Performance of the algorithms with the data set Contraceptive	193

B.14 Performance of the algorithms with the data set Leukemia discretized with equal frequency	194
B.15 Performance of the algorithms with the data set Leukemia discretized with equal width	195
B.16 Performance of the algorithms with the data set Lymphoma dis- cretized with equal frequency	196

List of Tables

2.1	An example of a transactional data set	9
2.2	Summary of algorithms and sub-problems related to frequent pattern mining	18
3.1	Example of the expressiveness of disjunctive and conjunctive patterns in a fictitious data set	22
3.2	An example data set	24
3.3	Example of itemsets from the data set in Table 3.2	24
3.4	List of notations used in Algorithm 1 describing the TitanicOR algorithm	29
3.5	Candidate minimal generators obtained from singletons in Example 1	32
3.6	Candidate minimal generators of size two in Example 1	32
3.7	Candidate minimal generators of size three in Example 1	33
3.8	Parameters used to generate synthetic data sets for evaluating the performance of TitanicOR	34
3.9	Characteristics of real world data sets used for evaluating the performance of TitanicOR	37
3.10	Description of the configurations used for evaluating the impact of maximum support, minimum support, and maximum number of items per itemset thresholds in TitanicOR and BLOSOM	37
3.11	Average number of itemsets, candidates, and average time spent per candidate by each algorithm in Test 1	39
3.12	Average number of itemsets, candidates, and average time spent per candidate by each algorithm in Test 3	39
4.1	An example data set to illustrate concepts related to Disclosed	51
4.2	Examples of the application of functions α and β on sets of samples and features on Table 4.1a	52
4.3	Transposed conditional tables for sample sets from Table 4.1a	56
4.4	The steps of the execution of Disclosed on the data set of Table 4.1a	63
4.5	Description of discretized data sets used for assessing the performance of Disclosed	66
4.6	Classification of the data sets according to Disclosed’s performance	69
5.1	Example data set for quasi-CNF emerging patterns	88
5.2	Description of data sets used for evaluating the performance of QCEP	97
5.3	Characteristics of data set Colon for evaluating the performance of the QCEP method	98

5.4	Characteristics of data set ALL–AML for evaluating the performance of the QCEP method	99
5.5	Characteristics of data set Leukemia for evaluating the performance of the QCEP method	100
5.6	Characteristics of data set Lymphoma for evaluating the performance of the QCEP method	101
5.7	Success rate of QCEP algorithms	103
5.8	Size of the right border of data sets	108
6.1	Summary of the characteristics of data sets affecting the performance of our methods	123
A-1	Description of the sources of the data sets used to assess the performance of Disclosed	142
A-2	Characteristics of the ALL–AML data sets considering different minimum support thresholds	143
A-3	Characteristics of the Colon data set considering different minimum support thresholds	144
A-4	Characteristics of the Embryo data set considering different minimum support thresholds	145
A-5	Characteristics of the GDS963 data set considering different minimum support thresholds	146
A-6	Characteristics of the GDS2200 data set considering different minimum support thresholds	146
A-7	Characteristics of the GDS2250 data set considering different minimum support thresholds	147
A-8	Characteristics of the GDS2519 data set considering different minimum support thresholds	148
A-9	Characteristics of the GDS2545 data set considering different minimum support thresholds	148
A-10	Characteristics of the GDS2821 data set considering different minimum support thresholds	149
A-11	Characteristics of the GDS2941 data set considering different minimum support thresholds	149
A-12	Characteristics of the Leukemia data set considering different minimum support thresholds	150
A-13	Characteristics of the Lymphoma data set considering different minimum support thresholds	151
A-14	Characteristics of the Promoters data set considering different minimum support thresholds	152
A-15	Raw performance results for Disclosed with the data set GDS963 . . .	162
A-16	Raw performance results for Disclosed with the data set GDS2200 . .	163
A-17	Raw performance results for Disclosed with the data set GDS2545 . .	164
A-18	Raw performance results for Disclosed with the data set GDS2821 . .	165
A-19	Raw performance results for Disclosed with the data set GDS2941 . .	166
A-20	Raw performance results for Disclosed with the data set Leukemia . .	167
A-21	Raw performance results for Disclosed with the data set Embryo . . .	168
A-22	Raw performance results for Disclosed with the data set GDS2519 . .	169
A-23	Raw performance results for Disclosed with the data set Lymphoma .	170

A-24	Raw performance results for Disclosed with the data set Promoters	171
A-25	Raw performance results for Disclosed with the data set ALL-AML	172
A-26	Raw performance results for Disclosed with the data set Colon	173
A-27	Raw performance results for Disclosed with the data set GDS2250	174
B-1	Performance of the algorithms in the experiments of Section 5.5 with data set ALL-AML discretized with equal frequency	175
B-2	Performance of the algorithms in the experiments of Section 5.5 with data set Colon discretized with equal width	177
B-3	Performance of the algorithms in the experiments of Section 5.5 with data set Colon discretized with equal frequency	179
B-4	Performance of the algorithms in the experiments of Section 5.5 with the data set Contraceptive	181
B-5	Performance of the algorithms in the experiments of Section 5.5 with data set ALL-AML discretized with equal width	181
B-6	Performance of the algorithms in the experiments of Section 5.5 with data set Leukemia discretized with equal frequency	183
B-7	Performance of the algorithms in the experiments of Section 5.5 with data set Leukemia discretized with equal width	185
B-8	Performance of the algorithms in the experiments of Section 5.5 with data set Lymphoma discretized with equal frequency	187
B-9	Performance of the algorithms in the experiments of Section 5.5 with data set Lymphoma discretized with equal width	189

List of Algorithms

1	TitanicOR	29
-	Function TitanicORGen	30
2	The algorithm Disclosed	60
-	Procedure $\text{Traverse}(X, s, TT _X)$ — Disclosed’s search space traverse algorithm	61
3	An algorithm for mining jumping QCEP borders	93

Abstract

Frequent itemset mining is one of the most studied problems in data mining. Since Agrawal et al. (1993) introduced the problem, several advances both theoretical and practical have been achieved. In spite of that, there are still many unresolved issues to be tackled before frequent pattern mining can be claimed a cornerstone approach in data mining (Han et al., 2007). Here, we investigate issues related to: (1) the (un)suitability of frequent itemset mining algorithms to identify patterns in biomedical data sets; and (2) the limited expressiveness of such patterns, since, in its vast majority, frequent itemsets are exclusively conjunctions.

Our ultimate goal in this thesis is to improve methods for frequent pattern mining in such a way that they provide alternative insightful solutions for mining biomedical data sets. Specifically, we provide efficient tools for mining disjunctive patterns in biomedical data sets. We tackle the problem of mining disjunctive patterns through three different fronts: (1) disjunctive minimal generators; (2) disjunctive closed patterns; and (3) quasi-CNF emerging patterns. We then propose three different algorithms, one for each task above: *TitanicOR*, *Disclosed*, and *QCEP*. While the first two aim for more descriptive patterns, the third is a more predictive.

These algorithms are proposed as an attempt to cover different sources of data sets coming from biomedical researches. *TitanicOR* is more suitable to identify patterns in data sets containing physiological, biochemical, or medical record information. *Disclosed* was designed to exploit the characteristics of microarray gene expression data sets, which usually contains many features, but only few samples. Finally, *QCEP* is the only algorithm to consider data sets with class label information. We conducted experiments with both synthetic and real world data sets to assess the performance of our algorithms. Our experiments show that our algorithms overcame the state of the art algorithms in each of those categories of patterns.

