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

List of Figures

List of Tables

List of Algorithms

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.