Assessing a feature's trustworthiness and two approaches to feature selection

Franco Alessandro Ubaudi

THESIS

Submitted to the Faculty of Engineering and IT
University of Technology Sydney (UTS)
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computing Sciences

2011

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature sources used are indicated in the thesis.

Signature of Student

_____

Acknowledgements

I am in debt to many people—for without their help I would never have made it. First of all I wish to thank Dr. Paul J. Kennedy from the University of Technology Sydney (UTS), for you were my greatest source of encouragement and direction. Among other things Paul, you provided me with a high standard to shoot-for and although it was tuff going at times, your standards enabled me to grow even further in my abilities and I am truly grateful for that. I am also grateful for your *immense* patience and *continuous* words of encouragement. You are a great supervisor!

I owe a debt of gratitude to Adjunct Professor Daniel Catchpoole from The Children's Hospital at Westmead (CHW) for his encouragement, support and direction. In particular I am very grateful for the times you called me and provided timely words of encouragement. How many supervisors are there who chase up their students; but fortunately I had such a supervisor. There were times you even drove me home after a meeting, just so we could chat. Thanks also to Nicholas Ho at CHW, for helping me get my head-around *R*, for your practical support in building heat maps and using GSEA.

Thanks to Professor Simeon J. Simoff, head of the school of computing at the University of Western Sydney (UWS), for your support and encouragement. In particular you have a great ability to see the big picture and that often provided the needed focus on where to head, which was always valuable since it is so easy to be short-sighted when there is plenty of detailed work to do.

Thanks also to Jenny and Lyn for your practical support over the years. In particular Jenny, I wish to thank you for making a big deal when I completed my Masters degree and the surprise party, it was amazing. Thank you Lyn, for whenever I was down and needed someone to believe in me, you were always what I needed. I also owe a debt of thanks to Uncle Dick from the "Land of the 10 10". I have known two incredible men in my life and you are one of them. In addition to being a true man of God, you always chase me up and show interest in where I am at and you always provide words of encouragement. Your are an amazing man and I hope that some of you will rub off onto me.

I am also in debt to Blair, Kathryn and Nicola, for each of you were extremely understanding, patient and supportive to me during this long haul. I am so privileged in being your dad. I am proud of you guys and grateful for our friendship and the fun we have together. You are the best kids a dad

could ever hope for!

I wish to pay tribute to my parents Giovanni and Enza, for they sacrificed a great deal so I could have opportunities they never had. The memory of you two will always be with me, for you were great parents and I lacked *nothing* important. I will never forget you mum and dad, or what you taught me. May I always make both of you proud of me.

Thanks also to my work colleagues: Greg Edwards, Peter Stephanou, Melanie Wilmot, Aaron Ashe, Karen Parker, Peter Gaudron and the rest of you in Distribution Planning. For you guys were *always* flexible, supportive and a pleasure to work with.

But most of all I am entirely in debt to my God and savior, for he provided me with my abilities and the opportunities to use them. He opened and closed doors at the right time in my life and although I often failed to understand the timeliness of it all—looking back now, I realize more than ever that your timing and provision is always perfect. Among many other things, he provided me the opportunity to undertake this research and the perfect job at Ausgrid (formally Energy Australia), which in particular provided the needed flexibility and money to pay the bills. Thanks also for your word: for it provides me with a light showing the way and a secure foundation that will *never* fail me.

> Your word is a lamp to my feet and a light for my path.
> (Psalms 119:105 NIV)
>
> Therefore everyone who hears these words of mine and puts them into practice is like a wise man who built his house on the rock. The rain came down, the streams rose, and the winds blew and beat against that house; yet it did not fall, because it had its foundation on the rock. (Matthew 7:24-25 NIV)

Contents

List of Figures

List of Tables

Abstract

Improvements in technology have led to a relentless deluge of information that current data mining approaches have trouble dealing with. An extreme example of this is a problem domain that is referred to as"non-classical". Non-classical problems fail to fulfill the requirements of statistical theory: that the number of instances in the sample set be much greater than the number of dimensions. Non-classical problems are mainly characterized by many dimensions (or features) and few noise-affected samples.

Microarray technology provides one source of non-classical problems, which typically produces data sets with a dimensionality exceeding ten thousand and containing just a few hundred instances. A risk with such a data set is building a model that is significantly influenced by coincidental correlations between the inputs (or the model's features) and the output. A classical strategy for managing this risk is reducing the dimensionality without significantly affecting the correlation between the remaining features and the model's output. However this strategy does not explicitly consider the impact

of poor data quality (or noise) and having few data samples.

In order to actively manage noise—a feature selection strategy is needed that not only considers the correlation between the features and the output, but also the quality of the features. It is proposed that feature quality, or simply the feature's "trustworthiness", should be incorporated within feature selection. As the trustworthiness of a feature increases, it is expected that the ability to accurately extract the underlying structure of the data will also increase. Another characteristic of non-classical problems is significant feature redundancy (where information provided within one dimension is also present in one or other dimensions). This research postulates that the use of feature trustworthiness and redundancy provides an opportunity to actively reduce the noise associated with the selected feature set, while still finding features that are well correlated with the model's output.

Two fundamental contributions are provided by this thesis: the notion of feature "trustworthiness" and how trustworthiness can be integrated within feature selection. Trustworthiness provides a flexible approach for evaluating the quality of a feature's sample data and in certain cases, the quality of the test data. This flexibility encourages the use of prior knowledge about the specific problem and in particular, how the quality of the data is best estimated. Traditionally feature selection implicitly assumes that every instance of data, supplied by preprocessing, has the same quality. Trustworthiness

also provides an opportunity for incorporating a measure of the changes applied to the data set as a result of data cleaning.

Using an area of computational learning, a theoretical justification was constructed that showed the difficulty of building an accurate model for a non-classical problem. The justification showed how a modest data quality problem can result in insufficient sample data to permit successful learning. It also showed how selecting less noisy data, or sufficiently trustworthy features, can enable successful learning using the available data points.

This thesis presents two methodologies that incorporate a measure of data quality within feature selection: one methodology only uses training data, while the other also incorporates test data while evaluating feature trustworthiness. The two methodologies are contrasted with each other and with a traditional feature selection methodology, which does not consider data quality.

A number of data sets were used to test these methodologies, with the main data sets being: synthetic data, childhood leukaemia and chronic fatigue syndrome. In most cases the three feature selection methodologies achieved similar accuracy however there were clear differences in the features selected by each. Using heat maps to visualize the clarity of the separation of the class labels by the selected features—showed dramatic differences. The two methodologies that incorporate trustworthiness provided a clearer

separation, while the traditional methodology was substantially inferior and appeared to be heavily influenced by artifacts. Using Gene Set Enrichment Analysis (GSEA), a widely used resource for evaluating the biological meaningfulness of gene sets (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov, 2005), showed that the two proposed methodologies selected genes that were more biologically meaningful than those selected by a traditional feature selection methodology.

The experiments also evaluated the sensitivity of trustworthiness to differences in the data set. By evaluating the trustworthiness of every feature, it was shown that considerable changes occurred across data folds. This result agrees with findings in the literature, such as (Ein-Dor, Kela, Getz, Givol, and Domany, 2005) and provides one explanation for the difficulty of modeling non-classical problems.

Contributions to knowledge

1. **Two approaches for estimating data quality**

   This thesis describes two approaches for estimating the quality of a feature's data set, one using a measure of signal-to-noise and the other, only a measure of noise. The latter approach permits the use of test data within the estimate of a feature's utility. In calculating the quality of a feature's data set, both approaches use prior knowledge about the data set.

2. **Definition and measure of feature trustworthiness**

   This thesis defines "trustworthiness" as an overall measure of the quality of a feature. Trustworthiness is required since each item of data associated with a feature has a unique level of quality. As a feature's trustworthiness increases, its utility approaches the estimated information it provides.

3. **Definition and measure of feature utility**

   Feature selection methodologies typically evaluate a feature's usefulness according to the feature's inherent information. However, noise may falsely increase its information; hence reduce the feature's trustworthiness. This thesis describes a measure of a feature's usefulness called "feature utility", which is a function of information provided by the feature and the noise present within its data set.

4. **Two feature selection methodologies based on feature utility**

   This thesis develops and evaluates two feature selection methodologies based on feature utility. The methodologies, "Feature Utility Ranking 1" and "Feature Utility Ranking 2", use respectively, a signal-to-noise ratio and a noise-only measure of quality.

   The methodologies were evaluated using computational learning theory, a series of synthetic data experiments and three biomedical data sets, which are characterized by high dimensionality, few samples and significant noise. The biomedical data experiments consist of leukaemia cell type classification, leukaemia treatment outcome prediction and chronic fatigue syndrome prediction.

5. **Mechanism for using data quality information collected during data cleaning**

   A mechanism for merging data cleaning outcomes within feature selection is explored. This approach allows the degree of change imposed by data cleaning to be incorporated in a feature's utility.

Publications

The work in this thesis has been published in:

Paul J. Kennedy, Simeon J. Simoff, Daniel R. Catchpoole, David B. Skillicorn, **Franco Ubaudi**, and Ahmad Al-Oqaily. Integrative Visual Data Mining of Biomedical Data: Investigating Cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. In Simoff Simeon J., Michael H. Bohlen, and Mazeika Arturas, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, volume 4404 of Lecture Notes in Computer Science, pages 367-388. LNCS, Springer, Berlin Heidelberg, 2008.

Franco A. Ubaudi, Paul J. Kennedy, Daniel R. Catchpoole, Dachuan Guo, and Simoff Simeon J. Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking. In L. Cao, P. S. Yu, C. Zhang, and H. Zhang, editors, *Data Mining for Business Applications*, pages 159-168. Springer, 2009.