UNIVERSITY OF TECHNOLOGY SYDNEY

Faculty of Engineering and Information Technology

From Ambiguity and Sensitivity to Transparency and Contextuality

– A Research Journey to Explore Error-Sensitive Value Patterns in Data Classification

A thesis submitted

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

William Wu

Sydney, Australia

2019

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 a part of requirements for other degree except as fully acknowledged within the text.

I also certify that the 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 used are indicated in the thesis.

This research is supported by an Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Signature of Candidate

ABSTRACT

"*Error is not a fault of our knowledge, but a mistake of our judgment giving assent to that which is not true*". This statement by John Locke, an English philosopher and medical researcher in the 1680s, is still relevant today, and this scope of error can be expanded from knowledge and judgement to result and process in terms of data analysis, to treat errors as a part of the knowledge to learn from rather than to simply eliminate.

In the research area of data mining and classification, errors are inevitable due to various factors such as sampling and computation restriction, and measurement and assumption limitations. To address this issue, one approach is to tackle errors head-on, to focus on refining the mining and classification processes by way of theory and algorithm enhancement to reduce errors, and it has been favored by researchers because the research results can be verified directly and clearly. Another approach is to focus on the examination of errors together with the data closely to explore the further understanding of different aspects of the data, especially on attributes and value patterns which may be more sensitive to errors to help identify and reduce errors in a retrospective and indirect way.

This research has taken up the latter and less favorable approach to learn from errors rather than simply eliminating them, to examine the potential correlation between the classification results and the specific characteristics of attributes and value patterns, such as value pattern ambiguity, error risk sensitivity and multi-factor contextuality, to help

enhance understanding of the errors and data in terms of correlation and context between various data elements for the goal of knowledge discovery as well as error investigation and reduction, not just for researchers, but more importantly, for the stakeholders of the data.

This research can be considered a four-stage journey to explore the ambiguity, sensitivity, transparency and contextuality aspects of value patterns from a philosophical and practical perspective, and the research work conducted in each stage of the journey is accompanied by the development of a new error pattern evaluation model to verify the results in a progressive and systematic way.

It is all about exploring and gaining further understanding on errors and data from different perspectives and sharing the developments and findings with the aim of generating more interest and motivation for further research into data and correlated factors, internally and externally, transparently and contextually, for the benefit of knowledge discovery.

ACKNOWLEDGEMENTS

I would like to thank the supervisors who guided me through this long and grueling course of study and development. Prof. Chengqi Zhang was my principal supervisor during my PhD candidature and I am very grateful for his insightful guidance, valuable advice and ongoing support. Dr Shichao Zhang, Dr Peng Zhang, Dr Jia Wu and Dr Jing Jiang all helped with many data exploration and risk investigation technique discussions at different stages of my study and research development, and their assistance and encouragement have been vital and greatly appreciated.

I must also give my sincere thanks to Prof. Judy Kay, A. Prof. James Athanasou and Ms. Michele Mooney as my PhD study and research development would not have been possible without the caring support and reference provided by Prof. Kay and A. Prof. Athanasou, and I would also like to thank Ms. Michele Mooney wholeheartedly for her thorough proofreading and constructive criticism.

Furthermore, I would like to express my immense gratitude to my family; any extra words of thanks would not be meaningful except my actions in resuming my share of the household duties, which include, but are not limited to, cleaning the kitchen and cooking the dinner.

It is in this context of acknowledgement and gratitude to colleagues and families, the realization of the importance of the context of our association to each other and the

impact of our lives to and between our environment, time and space and beyond becomes real.

VITA

Master of Education in Higher and Professional Education, University of Technology Sydney, 2008

Master of Information Technology, University of Sydney, 2003

Bachelor of Computing Science, University of Technology Sydney, 1997

PUBLICATIONS

W. Wu and S. Zhang. "Evaluation of error-sensitive attributes." Pacific-Asia Conference on Knowledge Discovery and Data Mining 2013, International Workshops on Trends and Applications in Knowledge Discovery and Data Mining, pp. 283-294. Springer, Berlin, Heidelberg. 2013.

W. Wu. "Exploring error-sensitive attributes and branches of decision trees." Ninth International Conference on Natural Computation (ICNC) 2013, pp. 929-934. IEEE. 2013.

W. Wu. "Identify error-sensitive patterns by decision tree." In: Perner P. (eds) Advances in Data Mining: Applications and Theoretical Aspects. ICDM 2015. Lecture Notes in Computer Science, vol 9165. Springer, Cham. 2015.

W. Wu. "Weakly Supervised Learning by a Confusion Matrix of Contexts." In: U. L., Lauw H. (eds) Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2019. Lecture Notes in Computer Science, vol 11607. Springer, Cham. 2019.

Paper Accepted for Presentation and Publication

W. Wu. "An Exploration into the Contexts of Errors and Value Patterns in Data Classification." Paper accepted for presentation in the 19th Industrial Conference on Data Mining ICDM 2019.

PREFACE

In the ancient parable of the blind men and an elephant, the six blind men are not labelled as liars because their false perceptions about the elephant are caused by their individual limitations and the lack of collaboration. On the other hand, each misunderstanding by these six blind men individually can be considered as one incomplete piece of a jigsaw and each can still contribute to a fuller understanding partially when properly connected. This is a classical reminder about the validity and importance of context consideration in these modern times with the latest technologies, to demonstrate that errors and misunderstanding can also contribute to our knowledge discovery and become part of a better understanding when they are explored and analyzed collaboratively and systematically, and with consideration of contexts in terms of external circumstantial factors and internal correlational influences.

Since errors are inevitable in the field of data mining and data classification, instead of following the old adage "if you can't beat them join them", it may be more sensible to “make friends” with errors and work with them closely in order to understand them better, rather than treating them as the enemy by avoiding them or getting rid of them at all costs.

As stated by Winston Churchill, “We have no lasting friends, no lasting enemies, only lasting interests.” Accordingly, in relation to our ongoing and lasting interest in knowledge discovery and problem solving, if we can gain a better understanding of the causes and implications of and the correlation between errors, such an understanding may in turn help expand our knowledge on data as a whole and help us formulate a more effective and lasting problem solution.

So, errors may not sound so bad after all. Depending on the context and perspective, errors may become a source of knowledge about data, a source of income for data scientists, a source of inspiration for researchers, and in this research, they become a source of determination to embark on a research journey to analyze errors and value patterns from various perspectives, to examine and evaluate errors and value patterns visually and correlatively, systematically and contextually, to better understand errors and data.

Within its own context of omission and admission, delusion and inspiration, this research has now reached its concluding stage and has been summarized into this thesis. It is hoped that this conclusion can serve as an invitation for more debate on the perpetual and contextual values and dealings of classification errors, to inspire further exploration into the contexts associated with errors and value patterns from a different perspective, to explore their potential correlation and causation based on science and technology with logical and empirical evidence and also with the consideration of tangible and philosophical contexts. Otherwise, technology without philosophy may lead to obscured perplexity and unsustainability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES