

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

| | |
|--|-------------|
| ABSTRACT..... | III |
| ACKNOWLEDGEMENTS | V |
| PREFACE..... | IX |
| LIST OF FIGURES | XVI |
| LIST OF TABLES | XVII |
| 1 INTRODUCTION..... | 1 |
| 1.1 Research Kindled by Errors and Contexts | 1 |
| 1.2 Thesis Statement | 4 |
| 1.3 Research Methodology and Tools | 5 |
| 1.4 Thesis Structure | 6 |
| 2 AN OVERVIEW AND ROADMAP OF THIS RESEARCH JOURNEY..... | 7 |
| 2.1 The Semantic and Philosophical Background at a Glance | 7 |
| 2.2 A Journey to Explore and Complement Algorithm-Focused Research..... | 12 |
| 2.3 Ambiguous Value Range and Error-Sensitive Attribute Evaluation | 13 |
| 2.4 Progress to Track down the Weakest Tree Branches and Value Patterns | 14 |
| 2.5 Tagging and Mapping the Weakest Link and the Achilles Heel for Transparency .. | 15 |
| 2.6 Transforming Confusion Matrices to Matrices of Contexts for Context Construction and Exploration..... | 16 |

| | | |
|----------|--|-----------|
| 3 | EVALUATION OF AMBIGUOUS VALUE RANGES AND ERROR-SENSITIVE ATTRIBUTES | 19 |
| 3.1 | Key Concepts | 21 |
| 3.2 | Initial Assumptions | 23 |
| 3.3 | How the Evaluation Process Works..... | 25 |
| 3.4 | Experiments | 28 |
| 3.4.1 | The Pima Indians Diabetes Dataset..... | 29 |
| 3.4.2 | The Wisconsin Breast Cancer Dataset | 34 |
| 3.4.3 | The Ecoli Dataset | 39 |
| 3.4.4 | The Liver Disorders Dataset | 40 |
| 3.4.5 | The Page Blocks Dataset..... | 43 |
| 3.5 | Experiment Analysis and Discussion..... | 46 |
| 3.5.1 | Evaluating Ambiguous Value Range by Overlapping Minimum-Maximum..... | 46 |
| 3.5.2 | Ambiguous Value Range by Mean Value of Class Labels | 50 |
| 3.5.3 | Discussion on Ambiguous Value Range using Other Methods | 53 |
| 3.5.4 | Comparing Attribute-Error Counter against Gain Ratio and Information Gain 54 | |
| 3.5.5 | Mixed Results from Potential Error-Reduction Measure..... | 59 |
| 3.5.6 | Follow-up Tests on the Error-Sensitive Attribute Hypothesis | 61 |
| 3.5.7 | Comparison with Experiments by Other Researchers..... | 63 |
| 3.6 | Related Work | 66 |
| 3.6.1 | Study of Ambiguous Data..... | 66 |

| | | |
|----------|---|------------|
| 3.6.2 | Comparing Attribute Selection..... | 69 |
| 3.7 | Summary..... | 71 |
| 4 | EXPLORATION OF THE WEAKEST AND ERROR-SENSITIVE DECISION BRANCHES | 73 |
| 4.1 | Overview of Decision Trees and Error-Sensitive Pattern Identification | 73 |
| 4.2 | Exploring Decision Trees for Error-Sensitive Branch Evaluation | 77 |
| 4.3 | EXPERIMENTS | 80 |
| 4.3.1 | The Connectionist Bench (Sonar, Mines vs. Rocks) Sonar Dataset | 81 |
| 4.3.2 | The MAGIC Gamma Telescope Dataset | 84 |
| 4.3.3 | The Yeast Dataset..... | 87 |
| 4.3.4 | The Cardiotocography Dataset..... | 91 |
| 4.3.5 | The Glass Identification Dataset | 93 |
| 4.4 | EXPERIMENT ANALYSIS | 96 |
| 4.4.1 | Comparison Between the Gain Ratio and Information Gain Approach..... | 97 |
| 4.4.2 | Effectiveness Considerations | 99 |
| 4.5 | Related Work | 102 |
| 4.5.1 | Decision Tree Classification | 102 |
| 4.5.2 | Tree- and Branch-based Study and Development..... | 103 |
| 4.6 | Summary..... | 106 |
| 5 | ENUMERATION OF DECISION TREES FOR PATTERN ANALYSIS WITH TRANSPARENCY | 108 |
| 5.1 | Initial Ideas about Tree Enumeration for Error Evaluation | 109 |

| | | |
|----------|--|------------|
| 5.2 | Background and Assumptions | 110 |
| 5.3 | Enumeration Process of a Decision Tree | 112 |
| 5.4 | Experiments | 114 |
| 5.4.1 | Enumeration Reflects Decision Trees in a Concise and Effective Way | 114 |
| 5.4.2 | Node-Level Statistics Comparison and Error-Sensitive Pattern Identification 121 | |
| 5.5 | Experiment Analysis | 125 |
| 5.5.1 | Digitized Node IDs Facilitate Node-Level Analysis..... | 125 |
| 5.5.2 | Examining the Weakest Nodes and Error-Sensitive Value Patterns..... | 126 |
| 5.5.3 | Possible Contributions, Effectiveness and Weakness | 128 |
| 5.6 | Related Work | 131 |
| 5.6.1 | Study of Classification Model Digitization..... | 131 |
| 5.6.2 | A More Practical Digitization Example | 133 |
| 5.7 | Summary | 135 |
| 6 | TRANSFORMATION OF CONFUSION MATRICES FOR CONTEXTUAL DATA ANALYSIS | 137 |
| 6.1 | Introduction..... | 137 |
| 6.2 | The Reasons and Potential Benefits of Transforming a Confusion Matrix | 138 |
| 6.2.1 | Why Context is Important in Error and Data Analysis? | 138 |
| 6.2.2 | Why Transform a Confusion Matrix into a Matrix of Context? | 141 |
| 6.2.3 | Key Components of the Confusion Matrix | 142 |
| 6.2.4 | Potential Benefits of Confusion Matrix Transformation..... | 143 |

| | | |
|----------|---|------------|
| 6.3 | Context Construction with Confusion Matrix..... | 146 |
| 6.3.1 | Post-Classification Analysis and Data Normalization | 146 |
| 6.3.2 | Construction of Categorical, Incremental and Correlational Context..... | 147 |
| 6.4 | Experiments and Analysis..... | 149 |
| 6.4.1 | Utilization of Existing Classification Models and Results | 149 |
| 6.4.2 | Constructing Contexts for a Lead Attribute in Pima Indians Diabetes Dataset | 150 |
| 6.4.3 | Constructing Contexts for Other Lead Attributes in the Pima Indians Dataset | 153 |
| 6.4.4 | Constructing Contexts for Other Datasets..... | 156 |
| 6.5 | Discussion of Potential Issues..... | 164 |
| 6.6 | Related Work | 167 |
| 6.6.1 | Study of Confusion Matrix and Contextual Analysis | 167 |
| 6.6.2 | The Adoption of Anna Karenina Principle in Other Studies | 171 |
| 6.7 | Summary | 173 |
| 7 | CONCLUSION AND FUTURE WORK | 175 |
| 7.1 | Summary of the Journey and Findings | 175 |
| 7.2 | Contributions..... | 180 |
| 7.3 | Future Study..... | 181 |
| | REFERENCES..... | 183 |

LIST OF FIGURES

| | |
|---|-----|
| Figure 2.1 – Examples of distinctive and overlapping value ranges | 14 |
| Figure 2.2 – Finding the weakest tree branches | 15 |
| Figure 3.1 – Two attribute value range situations | 23 |
| Figure 3.2 - Attribute and Error Evaluation as phase III of the classification process | 25 |
| Figure 3.3 – Three value range examples..... | 28 |
| Figure 3.4 - A typical feature ranking and selection model for data classification | 71 |
| Figure 4.1 - A sample tree with each branch showing its predicted error-rate..... | 76 |
| Figure 4.2 – Tree Branch Exploration is added to phase-III of the classification process | 78 |
| Figure 4.3 - Decision tree from the Connectionist Bench sonar dataset..... | 81 |
| Figure 4.4 - Decision tree from the MAGIC Gamma Telescope dataset..... | 85 |
| Figure 4.5 - Decision tree from the Yeast dataset | 88 |
| Figure 4.6 – Greatly simplified decision tree after removing highly error-sensitive attributes | 90 |
| Figure 4.7 - Decision tree from the Cardiotocography dataset..... | 91 |
| Figure 4.8 - Decision tree from the Glass ID dataset | 94 |
| Figure 5.1 - A decision tree example..... | 109 |
| Figure 5.2 – Pima Indians diabetes dataset’s decision tree model..... | 116 |
| Figure 5.3 - Wisconsin cancer dataset’s decision tree model | 117 |
| Figure 5.4 - Liver disorders dataset’s decision tree model | 119 |
| Figure 5.5 - Page Blocks dataset’s decision tree model | 120 |
| Figure 6.1 - Quarterly cash flow earnings ratio comparison between Enron and its industry model | 139 |
| Figure 6.2 - Enron's Year 2000 Reported Revenue vs. Similarly Sized Companies | 140 |
| Figure 6.3 - Class distribution of individual attributes in Pima Indians diabetes dataset | 145 |
| Figure 6.4 – Context construction using a confusion matrix during post-classification analysis..... | 146 |

LIST OF TABLES

| | |
|--|----|
| Table 2-1 – Example of decision tree and its digitized node IDs | 16 |
| Table 2-2 – Example of contextual comparison using a confusion matrix..... | 17 |
| Table 3-1 – Pima dataset’s initial classification result by J48 in WEKA | 29 |
| Table 3-2 - Two methods of calculation for the Pima dataset’s ambiguous value range | 30 |
| Table 3-3 – The Pima dataset’s attribute-error counter, gain ratio and information gain ranking lists | 33 |
| Table 3-4 - Comparing Pre- vs. Post- removal of error-sensitive attributes in the Pima dataset..... | 34 |
| Table 3-5 - Wisconsin dataset’s initial classification result | 34 |
| Table 3-6 - Two methods of calculation for the Wisconsin dataset’s ambiguous value range..... | 35 |
| Table 3-7 - Wisconsin dataset’s attribute-error counter, gain ratio and information gain ranking lists | 37 |
| Table 3-8 - Pre- vs. Post- removal of error-sensitive attributes in the Wisconsin dataset | 38 |
| Table 3-9 - Ecoli dataset’s attribute-error counter, gain ratio and information gain ranking lists..... | 39 |
| Table 3-10 - Pre- vs. Post- removal of error-sensitive attributes in the Ecoli dataset..... | 40 |
| Table 3-11 - Liver Disorders dataset's attribute-error counter, gain ratio and information gain ranking lists | 41 |
| Table 3-12 - Pre- vs. Post- removal of error-sensitive attributes in the Liver Disorders dataset..... | 42 |
| Table 3-13 - Page Blocks dataset's attribute-error counter, gain ratio and information gain ranking lists | 43 |
| Table 3-14 - Pre- vs. Post- removal of error-sensitive attributes in Page Blocks dataset | 44 |
| Table 3-15 - Pima diabetes dataset’s ambiguous value range by overlapping minimum and maximum of TN & TP attribute values | 47 |
| Table 3-16 - Wisconsin cancer dataset’s ambiguous value range by overlapping minimum and maximum of TN & TP attribute values..... | 48 |
| Table 3-17 - Outlier values amongst negative/positive samples in the Wisconsin dataset..... | 49 |

| | |
|--|-----|
| Table 3-18 – Pima diabetes dataset’s ambiguous value range by mean value of class label..... | 51 |
| Table 3-19 - Wisconsin dataset’s ambiguous value range by mean value of class label..... | 52 |
| Table 3-20 – Comparison of the top-3 attributes ranked by attribute-error counter, gain ratio and information gain algorithm in the Pima diabetes dataset | 55 |
| Table 3-21 - Top-4 attributes ranked by attribute-error counter, gain ratio and information gain algorithm in the Wisconsin dataset..... | 55 |
| Table 3-22 - Comparison of attribute-error count ratio between the Pima and Wisconsin datasets | 57 |
| Table 4-1 – Exploration of error-sensitive branches in the Sonar dataset | 82 |
| Table 4-2 - Pre- vs. Post- removal of error-sensitive attributes in the Sonar dataset..... | 84 |
| Table 4-3 - Exploration of error-sensitive branches in the Gamma dataset..... | 86 |
| Table 4-4 - Pre- vs. Post- removal of error-sensitive attributes in the Gamma dataset | 87 |
| Table 4-5 – Exploration of error-sensitive branches in the Yeast dataset | 89 |
| Table 4-6 - Pre- vs. Post- removal of error-sensitive attributes in the Yeast dataset..... | 90 |
| Table 4-7 - Exploration of error-sensitive branches in the Cardiotocography dataset | 92 |
| Table 4-8 - Pre- vs. Post-removal of error-sensitive attributes in the Cardiotocography dataset | 93 |
| Table 4-9 - Exploration of error-sensitive branches in the Glass ID dataset | 95 |
| Table 4-10 - Pre- vs. Post-removal of error-sensitive attributes in the Glass ID dataset..... | 96 |
| Table 4-11 – Comparing error rate and error count for three branches in the Gamma dataset..... | 98 |
| Table 5-1 - Ecoli dataset's decision tree in digitized form..... | 115 |
| Table 5-2 - Pima Indians dataset's decision tree represented by enumerated leaf-node IDs | 116 |
| Table 5-3 - Wisconsin dataset's decision tree represented by enumerated leaf-node IDs | 118 |
| Table 5-4 - Liver disorders dataset's decision tree represented by enumerated leaf-node IDs..... | 119 |
| Table 5-5 - Page Blocks dataset's decision tree represented by enumerated leaf-node IDs | 120 |
| Table 5-6 - The weakest nodes' attributes & values vs. the most error-sensitive attributes | 122 |
| Table 5-7 - Three-way comparison - original vs weakest node vs most error-sensitive attribute | 124 |

| | |
|---|-----|
| Table 5-8 – Basic process in digitizing artificial neural networks | 133 |
| Table 5-9 – Digital feedback by latches as a form of back-propagation | 134 |
| Table 6-1 – A Confusion matrix of contexts constructed for the Page Blocks dataset..... | 142 |
| Table 6-2 - Classification result for Pima diabetes dataset and options for lead attributes | 149 |
| Table 6-3 - Establish the categorical and incremental context for <i>plas</i> as the lead attribute | 150 |
| Table 6-4 - Matrix of incremental, correlational and categorical context for <i>plas</i> | 152 |
| Table 6-5 - Matrix of incremental, correlational and categorical context for <i>mass</i> | 154 |
| Table 6-6 – Plotting the matrix of incremental, correlational and categorical context..... | 155 |
| Table 6-7 - Classification result for Page Blocks dataset and options for lead attributes..... | 157 |
| Table 6-8 - Establishing categorical and incremental context for <i>length</i> as the lead attribute | 158 |
| Table 6-9 - Matrix of incremental, correlational and categorical context for <i>length</i> as the lead attribute... | 159 |
| Table 6-10 - Establishing categorical and incremental context for <i>Normal_Nucleoli</i> as the lead attribute. | 162 |
| Table 6-11 - Matrix of incremental, correlational and categorical context for <i>Normal_Nucleoli</i> | 163 |
| Table 6-12 –Confusion matrix for two-class classification..... | 168 |