

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

<b>ABSTRACT.....</b>	<b>III</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>V</b>
<b>PREFACE.....</b>	<b>IX</b>
<b>LIST OF FIGURES .....</b>	<b>XVI</b>
<b>LIST OF TABLES .....</b>	<b>XVII</b>
<b>1 INTRODUCTION.....</b>	<b>1</b>
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
<b>2 AN OVERVIEW AND ROADMAP OF THIS RESEARCH JOURNEY .....</b>	<b>7</b>
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

<b>3</b>	<b>EVALUATION OF AMBIGUOUS VALUE RANGES AND ERROR-SENSITIVE ATTRIBUTES .....</b>	<b>19</b>
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
<b>4</b>	<b>EXPLORATION OF THE WEAKEST AND ERROR-SENSITIVE DECISION BRANCHES .....</b>	<b>73</b>
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
<b>5</b>	<b>ENUMERATION OF DECISION TREES FOR PATTERN ANALYSIS WITH TRANSPARENCY .....</b>	<b>108</b>
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
<b>6</b>	<b>TRANSFORMATION OF CONFUSION MATRICES FOR CONTEXTUAL DATA ANALYSIS.....</b>	<b>137</b>
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
<b>7</b>	<b>CONCLUSION AND FUTURE WORK .....</b>	<b>175</b>
7.1	Summary of the Journey and Findings .....	175
7.2	Contributions.....	180
7.3	Future Study.....	181
	<b>REFERENCES.....</b>	<b>183</b>

## 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

## CHAPTER 1

### Introduction

#### 1.1 Research Kindled by Errors and Contexts

The Pima Indians diabetes dataset hosted by the UCI Machine Learning Repository [BL13] is widely used in data classification experiments, which may be partially because the original owner, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the United States, is an organization with an impeccable reputation “to support research, training, and communication with the public in the topic areas of "diabetes and other endocrine and metabolic diseases” ([www.niddk.nih.gov](http://www.niddk.nih.gov)). Even though the UCI’s Pima diabetes dataset of 768 records is small compared to the vast amount of data on diabetes and other metabolic diseases collected by NIDDK for research, this dataset with 500 negative records and 268 positive records is an effective and reliable example for measuring and comparing the performance of a variety of classification algorithms in a general sense. However, extra caution is required if this dataset is used for research and experiments on a specific health and diabetes research topic.

One key reason for the extra caution concerns the need for the consideration of historical, geographical and economic context in relation to specific links between the Pima Indians from the Gila River Indian Community in Arizona and the diabetes epidemic. It was noted that diabetes was uncommon amongst the Pima Indians until the 1930s but became an epidemic during the 1950s and 1960s [Nar97] [Pil12] [SBR06]. Key factors

highlighted in the research include the fact that in the 1890s, the Pima Indians lost their access to water from the Gila River, adversely affecting their agricultural supply, the heavy and physical work associated with farming, and also the adaption of modern Western fast food and lifestyle. In comparison, another group of Pima Indians of the same tribe migrated to Mexico in the 1890s and continued their physical and traditional farm work and traditional diet, and have a significant lower rate of diabetes despite these two groups sharing a common genetic background [VWN05] [Pil12].

Such consideration and analysis within the wider context of the Pima Indians' diagnostic data has prompted new thoughts on a common approach of singling out a specific group of genes and focusing on research for genetic and medicine remedy, as suggested in Phihl's 2012 report, to generate more interest in and research on "natural and social environmental causes of diabetes" in addition to the usual targets such as diagnostic analysis and medical care development. This is because "Many pharmaceutical company's economic best interests lay within funding research that will provide them the opportunity to provide retrospective care" (Phihl, 2012, p.3), which may, intentionally or unintentionally, mislead the research.

These historical and environmental factors can be considered as external contexts when they are not included as a part of a dataset for research; meanwhile, some forms of internal and implicit contexts may also exist inside the dataset in a less observable way. Referring to the study of diagnostic data for diabetes, examples of internal context may include diagnostic information about lungs, stomach and kidneys, as these body organs are considered to be closely associated with diabetes according to the traditional Chinese

medicine (TCM) theory and because of their common metabolic functions with food and drink, urination and energy circulation [Cov01] [WWC13]. Other observational and contextual information may include the color and texture of the face and tongue, the odor of the breath and body, and the strength and rhythm of the pulse, all regarded as “a detailed, multi-system case history and [.....] this information [is supplemented] with observations that give information about the state of the patient’s health” (Covington, 2001, p.3). Based on such contexts, “TCM views the human body and its functioning in a holistic way ... looks at patterns of disharmony, which include all presenting signs and symptoms as well as patients’ emotional and psychological responses” (Covington, 2001, p.1). Therefore, instead of focusing on “lowering blood glucose levels” with “western medicine which usually contains a single active ingredient aiming for a specific mechanism” (Wang, Wang & Chan, 2013, p.5), herbal medicine “may contain various active ingredients targeting multiple mechanisms ... based on the holistic theory, which puts an emphasis on the integrated body” (Wang et al., 2013, p.5).

When information on internal factors and contexts is absent or is not presented in a perceivable manner, it can then become a task for data researchers to collect and collate additional details by collaborating with other domain experts, and to uncover more information by the further exploration and analysis of the existing data and the related factors. This study has taken on this data analytic task, and has developed a systematic process that can construct a dimensional contextual environment based on existing data and the associated classification process, to simulate the growth of key attributes and visualize the correlational effect between attributes and between classification categories within a dataset,

to help gain further understanding about the data and the interrelationship between value patterns and classification results.

Prior to this stage of context model construction based on the growth simulation of key attributes, there are three preparatory stages involved to provide the initial building blocks and platform. The first stage is identifying the most error-sensitive attributes as the key attributes by evaluating their ambiguous value ranges and related error statistics using decision trees; the second stage is identifying the most error-sensitive value patterns in the form of the weakest decision tree branch; and the third stage is digitizing a decision tree into a map of enumerated tree nodes with unique digital ID tags, which can help store and track classification statistics at individual nodes, making the evaluation process of error-sensitive value patterns more transparent and effective. More detail on these tasks is given in the next few chapters.

## **1.2 Thesis Statement**

This thesis suggests the identification and analysis of specific characteristics in data records, such as error-sensitive patterns of data attributes and values, can enhance the understanding of data and may, in turn, also help improve the detection and measurement of error for the underlying classification process in a retrospective and reversed-engineering way. In supporting this suggestion, four data evaluation models have been developed to identify and analyze error-sensitive value patterns in terms of ambiguity and sensitivity, transparency and contextuality, which is a shift from the usual focus on algorithm development to argue that errors can be part of the data, and the exploration and comprehension

of error-sensitive value patterns can potentially lead to a better understanding and knowledge discovery of the data.

### **1.3 Research Methodology and Tools**

The research work documented in this thesis contains both qualitative and quantitative study and includes theoretical and empirical analysis. Books and journal and conference papers identified in literature surveys are the primary sources of information on theories and algorithms. Creditable webpages are also viewed and quoted but not cited as a primary reference. The UCI Machine Learning Repository is the key source of experiment datasets.

WEKA is the primary data mining and analytic tool in this research work, and its J48 decision tree classifier has been used extensively in experiments to extract additional classification statistics for ambiguity evaluation and to enumerate a classification model for transparency analysis, and the required source-code level redevelopment is carried out under WEKA's open-source license. R is also used intermittently for result verification and comparison purposes. The Microsoft Access database is used for storing and collating the majority of the classification results, and Microsoft Excel is used for data manipulation and value plotting.

A number of customized functions have been added to WEKA by modifying part of its Java source code to handle some special requirements of this research, and a number of new Microsoft Access database modules have also been developed by coding in VBA to perform some specific data consolidation and integration tasks.



## 1.4 Thesis Structure

Chapter 1 introduces the motivation and key ideas in this study and provides an overview about the methodologies and tools used and the structure of this thesis.

Chapter 2 provides a broad review of the semantic and philosophical background of this study and outlines a roadmap of the study in the form of a research journey in four progressive stages - ambiguity, sensitivity, transparency and contextuality.

Chapter 3 provides a detailed account of the research work on the ambiguous value range analysis and error-sensitive attribute evaluation process and discusses experiment results with details on why some data are supportive and why other data are not.

Chapter 4 extends the error-sensitive attribute analysis and discussion to cover a cluster of related attributes and value ranges in a decision tree, to develop an error-sensitive decision branch evaluation process in helping to highlight the most risky and error-sensitive branch of a tree, which may, in turn, reflect the error-sensitivity level of the underlying section of the data. It is a part of the data understanding improvement process.

Chapter 5 takes error-sensitive pattern analysis into a digitized realm by enumerating a decision tree into a digitized map of node IDs, so the classification statistics of each leaf and node of a decision tree can be stored and retrieved uniquely for further analysis.

Chapter 6 builds a connection between the error and classification analysis by proposing a context construction model to help improve our understanding by examining data from various dimensional perspectives, such as categorical, incremental and correlational.

Chapter 7 summarizes this study and outlines a future development plan.

## CHAPTER 2

### **An Overview and Roadmap of this Research Journey**

Does the saying “can’t see the forest for the trees” also apply to this data mining and classification research if the focus of a study is only on the classification algorithm rather than a more inclusive approach with consideration of other internal and external factors as well? If the continuous classification enhancements by individual algorithm developments are compared to sparkling glitter which can shine with creativity but may or may not provide realistic value, then the improved understanding of a specific dataset or a particular type of data by team collaboration and factor correlation can be compared to “claiming gold” which can deliver practical and genuine value in terms of knowledge discovery. To date, it appears that a higher priority is given to general algorithm development and expansion than to specific data pattern examination and integrated understanding, and out-of-context findings with a limited scope can sometimes be misleading, as described in Chapter 1, which leads to the necessity of background studies on ambiguity, errors and contexts as discussed in this research.

### **2.1 The Semantic and Philosophical Background at a Glance**

According to Oxford Dictionary, the word “ambiguity” could mean “a word or statement that can be understood in more than one way” or “the state of having more than one possible meaning”, and it could also mean something “that it is unclear or confusing” or “contains several different qualities or attitudes which do not fit well together” according to Collins

Dictionary. In essence, the word *ambiguity* indicates there is more than one option available and that the number of options and the exact meaning of each option can be open for interpretation, which can subsequently cause confusion and lead to misinterpretation in semantic terms.

When conducting a data mining and classification study, especially when dealing with binary classification scenarios, some specific value patterns can lead to positive results in one group of records but can also be associated with negative results in another group of records. When such ambiguous value patterns are included as a part of the classification criteria to evaluate all data records using the underlying algorithm, developing a systematic process to identify and analyze such ambiguous value patterns may help in a better understanding of the value pattern transition between result categories and associate errors. This can be considered a data-centric method of pattern analysis to enhance error-reduction measures rather than the typical algorithm-centric approach focusing on the details of the underlying classification algorithm. The evaluation of such a systematic model for value ambiguity analysis is the key topic in the next chapter.

Generally speaking, ambiguity is regarded as a problematic issue; however, there are also views defending the purpose and existence of ambiguity. Finn of MIT [Fin12] suggested “ambiguity makes a language more efficient rather than less so”, and “it is ‘cognitively cheaper’ to have the listener infer certain things from the context than to have the speaker spend time on longer and more complicated utterances.” (Finn, 2012). This sounds like a defense for ambiguity from the point of view of efficiency and economy by the provider, rather than from the point of view of clarity and certainty by the receiver. If no prior

knowledge or baseline understanding can be assumed, the pursuit of clarity and continual questioning can result in an ongoing cycle of queries and uncertainty, and searching for root cause after root cause with no end, similar to Plato's recollection of Socrates suggestion, that "the primary elements are 'unaccountable and unknowable, but perceivable'" ([en.wikipedia.org/wiki/Theaetetus\\_\(dialogue\)](http://en.wikipedia.org/wiki/Theaetetus_(dialogue))).

On the other hand, what is an error? According to Oxford Dictionary, an error is "a mistake, especially one that causes problems or affects the result of something", and it could also mean "something you have done which is considered to be incorrect or wrong, or which should not have been done" according to the Collins Dictionary. In these semantic terms, an error is not regarded as a physical object like a rock or a tangible attribute such as color and temperature that exists independently of human recognition or perception.

Superficially speaking, an error can be considered as a kind of discrepancy between one's expectation and his/her observation and perception, a kind of subjective conceptual judgment rather than an objective physical element. Philosophically speaking, errors are inevitable because the limited human knowledge and capabilities are still evolving in this changing world, therefore errors can become a source of curiosity and a driving force for further learning and knowledge discovery in order to understand and manage the errors and data better.

Knowledge itself can be regarded as a kind of conceptual judgment based on factual observation and theoretical understanding within a certain scope such as space and time, and because everything changes over time, knowledge itself also evolves. This is why we learn from errors, and in this regard, errors can also be considered as an alternate or

opposite kind of knowledge, which may explain why Socrates believed that errors cannot occur because of the limitations in the perception of seeing and believing [Rob50].

Therefore, a more reasonable approach may be to only identify ambiguity if possible, then to clarify the context around the ambiguity where practical, to make an attempt to gain more clarity and a better understanding in a Socratic way ([https://en.wikipedia.org/wiki/Socratic\\_method](https://en.wikipedia.org/wiki/Socratic_method)). If and once ambiguity can be clarified, it may provide a better context on the topic and therefore errors can be better analyzed, which may subsequently lead to better results and improved understanding. As suggested by Plutarch, “To make no mistakes is not in the power of man; but from their errors and mistakes, the wise and good learn wisdom for the future.” ([https://www.brainyquote.com/quotes/plutarch\\_389110](https://www.brainyquote.com/quotes/plutarch_389110))

One way to help analyze and clarify the patterns and potential causes of ambiguity and errors is to provide a certain level of transparency on how the result of a data mining and classification task is achieved based on the underlying datasets and how classification errors are measured and reported. According to the Oxford Dictionary, one definition of transparency is “the quality of something, such as a situation or an argument, that makes it easy to understand”; similarly, the Collins Dictionary defines that “The transparency of a process, situation, or statement is its quality of being easily understood or recognized, for example because there are no secrets connected with it, or because it is expressed in a clear way.” If clarity and transparency on classification paths and associated elements and rules can be provided, it may help outline the data patterns and processing steps involved in triggering the classification errors, to help understand the result and errors with more

timing and conditional details as a form of contextual information, to learn and discover from trial and error in an expressive and explainable way.

This philosophical and conciliatory approach appears to express an appreciation of errors and mistakes rather than a denunciation, which serves as a part of the inspiration that drives this study to explore the issues of ambiguity and error from a variety of aspects and contextual perspectives.

Then, what is context? Semantically speaking and according to the Oxford Dictionary, context means “the words that come just before and after a word, phrase or statement and help you to understand its meaning”, or, “the situation in which something happens and that helps you to understand it”; and the Collins Dictionary defines context as “The context of an idea or event is the general situation that relates to it, and which helps it to be understood.” More generally speaking, context can be described as “any information that can be used to characterize the situation of an entity” (Dey and Abowd, 1999), and such an entity can be any relevant person or object or place [DA99]. It can also be expanded to include an event or a process that reflects the change of situation [Kov14].

In essence, context can relate to all the factors that are relevant to the topic in focus, physical and conceptual, situational and historical, all dimensions being dynamically inclusive. Despite the fact that context is more than a semantic definition confined by text, it does have certain constraints and limitations, such as a starting point and an end point, and a point of reference as its own context.

## 2.2 A Journey to Explore and Complement Algorithm-Focused Research

One key focus of data mining and classification research has been on the development and enhancement of theories and algorithms, to explore and compare between classification frameworks, to find the model, such as Bayesian, neural networks, decision trees, support vector machines and ensemble methods coupled with bagging and boosting techniques, which has a superior classification performance and higher accuracy. As a result, data and experiments are primarily used as peripheral elements for the sake of evidence to prove the academic success of such development and enhancement, and the improvement of understanding about the underlying data and the correlation between components and factors appears to be secondary.

Taking a different and more data-oriented approach and to make value pattern exploration and knowledge discovery a primary objective, this data classification study develops a number of systematic processes to explore and identify specific value patterns related to classification errors in an effective and transparent way, to help visualize and evaluate the correlation between classification errors and value patterns in various aspects from different perspectives. While this study does not directly work on the processing logic of a specific classification algorithm, its findings on specific value patterns and the contextual environment can potentially and indirectly benefit further algorithm enhancement with new considerations on error reduction measures, attribute clustering correlation and other factors due to an improved understanding of the data. Essentially, this adopted approach can be considered a data-centric complement rather than a direct algorithm-based development in the pursuit of data mining and classification goals.

This study can be considered as a research journey with four stages of progressive development, from ambiguity to sensitivity to transparency and contextuality. A roadmap of this journey is outlined in the following four sections.

### **2.3 Ambiguous Value Range and Error-Sensitive Attribute Evaluation**

Accuracy and error rate are the key benchmarks in measuring the level of success of a classification task and its underlying algorithms, and there are various factors that can cause misclassification errors in a data classification task. External factors may include the quality of the data measurement tools, the condition of the measurement, the method and environment of the data collection and the timing, etc. Internal factors may include the method of record sampling and training, threshold criteria of the adopted logic and formulas, overfitting and under-fitting issues of the statistical and processing model, etc.

Instead of looking into specific individual factors one by one at this stage, this study has developed a systematic way to identify attributes and their value ranges which may be more susceptible to the impact of error factors within the value transition range from the negative category to the positive category in binary classification scenarios, to help generate a kind of transitional context between negative and positive data samples, to add to the understanding of classification errors which are associated with specific attributes and their specific value patterns by defining and exploring three proposed terms: ambiguous value range, attribute-error counter and error-sensitive attribute. Further details are given in Chapter 3, however a simple illustration of this idea is depicted in Figure 2.1.



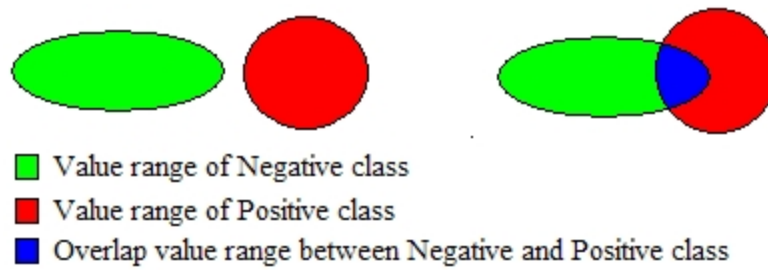


Figure 2.1 – Examples of distinctive and overlapping value ranges

## 2.4 Progress to Track down the Weakest Tree Branches and Value Patterns

A misclassified data sample with one or multiple attribute values inside their respective ambiguous value ranges may not necessarily be the sole reason for a misclassification error, but such ambiguity may have contributed to an increase in misclassification risk by influencing the underlying classification algorithm in a distracting way.

Contributing factors of classification error may include data quality issues such as the inclusion of noisy data, incorrect measurement of values, the mishandling of null values, overfitting and underfitting problems, etc. [Tay96] [Yoo89].

As an attempt to identify some of the factors and value patterns which may have caused or be associated with classification errors, the decision tree model has been adopted as an initial investigation platform for two key reasons. The first reason is because the efficiency and effectiveness of a decision model has been well researched [Qui86] [Sam93] [RM14], the second reason is because each tree branch represents a decision rule that lists the attributes and their value ranges involved in the classification process in a logical and graphical way. By evaluating and ranking classification errors in each decision tree branch, the potentially most error-prone classification rules could be identified in the form of the weakest branches of a decision tree. An evaluation process to identify the weakest tree

branch has subsequently been developed to expand the analytic scope of error-sensitivity from attribute-based to the inclusion of multiple attributes and their respective value ranges and hanging on the weakest tree branch when utilizing the decision tree classification algorithm, as shown in Figure 2.2.

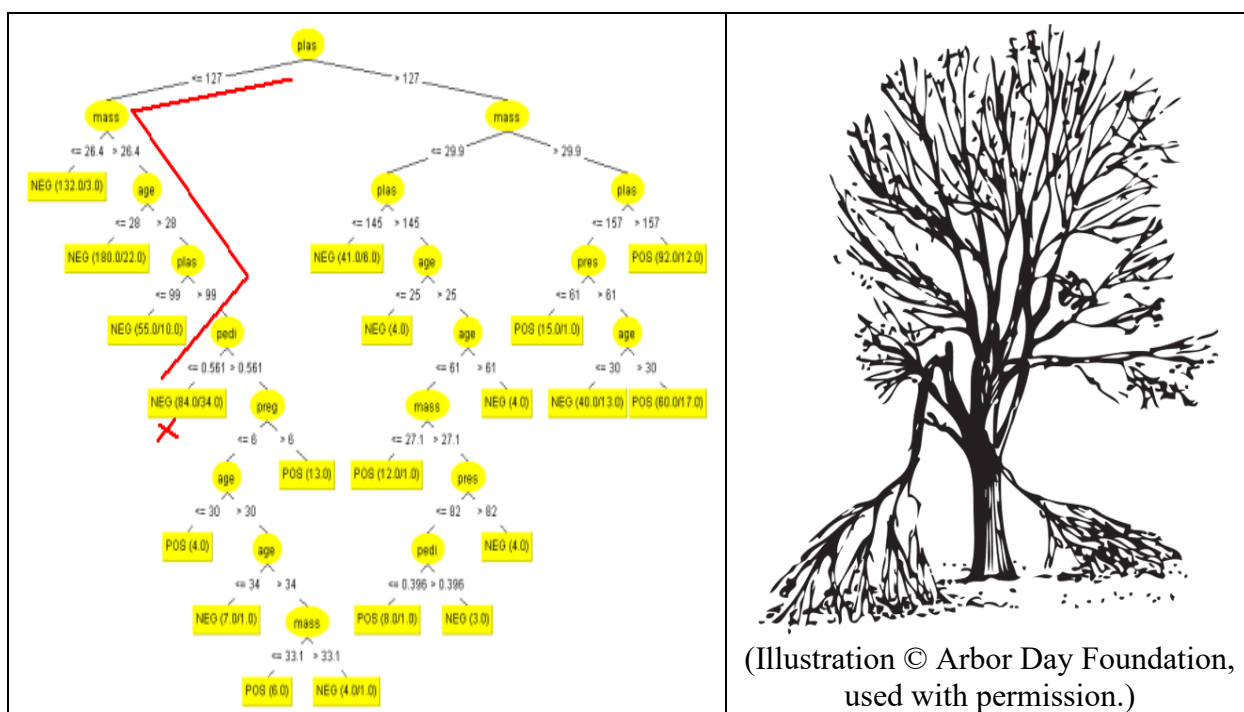


Figure 2.2 – Finding the weakest tree branches

## 2.5 Tagging and Mapping the Weakest Link and the Achilles Heel for Transparency

Data classification processes are sometimes considered black-box processes because the exact classification path for each data sample is often hidden and the classification result pops out of the black-box like magic. Attempts have been made to track through and map out the classification paths in a classification model in a more transparent way, such as a workflow reporting model and starting with the decision tree classification model and

making use of its “a-branch-represents-a-rule” advantage, a decision tree enumeration process has been developed.

This enumeration process digitizes each tree node into a unique numerical ID, enables classification statistics to be stored and accessible at individual node level, and helps identify the Achilles heel in the form of the weakest link in a classification tree by evaluating the node-level statistics in an effective and transparent way, as shown in Table 2-1.

Table 2-1 – Example of decision tree and its digitized node IDs

Pima diabetes dataset's decision tree model	Digitized Leaf-Node IDs
	Branch 1: 1.1.1 Branch 2: 1.1.2.1 Branch 3: 1.1.2.2.1 Branch 4: 1.1.2.2.2.1 Branch 5: 1.1.2.2.2.2.1.1 Branch 6: 1.1.2.2.2.2.1.2.1 Branch 7: 1.1.2.2.2.2.1.2.2.1 Branch 8: 1.1.2.2.2.2.1.2.2.2 Branch 9: 1.1.2.2.2.2.2 Branch 10: 1.2.1.1 Branch 11: 1.2.1.2.1 Branch 12: 1.2.1.2.2.1.1 ... Branch 18: 1.2.2.1.2.1 Branch 19: 1.2.2.1.2.2 Branch 20: 1.2.2.2

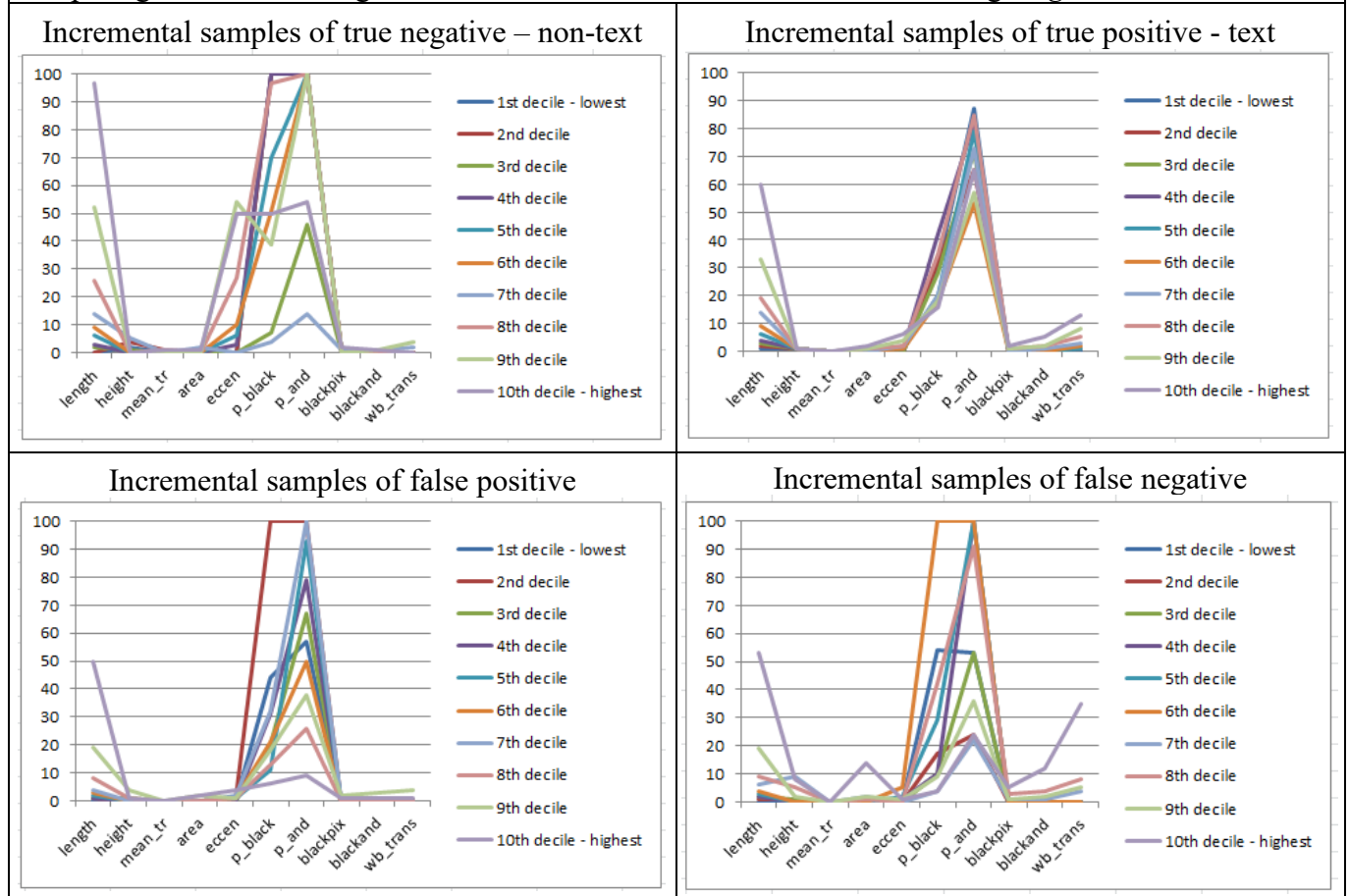
## 2.6 Transforming Confusion Matrices to Matrices of Contexts for Context Construction and Exploration

In a closer review of the progress and findings of the research journey so far, it has been recognized that while some specific ambiguous value ranges and highly error-sensitive attributes and value patterns have been identified and tested with supportive and encouraging

results, the true usefulness of these findings needs further insightful review and assessment by field domain experts and stakeholders with detailed knowledge, to avoid making incomplete and out-of-context claims.

Table 2-2 – Example of contextual comparison using a confusion matrix

Comparing a matrix of categorical, incremental and correlational context using *length* as the lead attribute



This leads to the next stage of the study on context matters in relation to data classification, and a context construction process has subsequently been developed to transform confusion matrices into matrices of contexts based on classification results and value patterns, to create and provide some forms of contextual environment as a part of the post-classification analysis, to help explore and evaluate potential relationships between value

patterns and classification results within their categorical, incremental and correlational context and in a simple, visual and systematic way, as shown in summary Table 2-2, and more details on the specific categorical, incremental and correlational context illustrated in this table are discussed in Section 6.4.4 with reference to Table 6-9.

Such multi-dimensional context construction is about helping to make value patterns and internal contexts more recognizable and interesting, and attempting to make the further understanding of data and knowledge discovery more perceivable and appealing.

After providing an initial discussion on ambiguity, error and context at a semantic and philosophical level, and outlining the roadmap for this research journey, the next few chapters report on the actual exploration and the experiments in more detail.

## CHAPTER 3

### **Evaluation of Ambiguous Value Ranges and Error-Sensitive Attributes**

This research journey starts with a study on ambiguity because ambiguity often leads to misunderstanding and errors. The term *error*, in the book “An introduction to error analysis” [Tay96], is described on page 3 of Chapter 1 as follows:

*“In science, the word error does not carry the usual connotations of the terms mistake or blunder. Error in a scientific measurement means the inevitable uncertainty that attends all measurements. As such, errors are not mistakes; you cannot eliminate them by being very careful. The best you can hope to do is to ensure that errors are as small as reasonably possible and to have a reliable estimate of how large they are.”*

This book also outlined some specific error categories, such as errors from direct measurements or indirect measurements, independent errors or correlated errors, random errors or systematic errors. In this regard, the term *error* is defined primarily in the context of “traditional” and natural science, such as physics, chemistry, biology and engineering, with tasks of measuring and processing raw and pre-processed data collected from the field materially and practically, directly and indirectly.

On the other hand, data mining and classification use new computer technology for information generation and knowledge formulation, to process and analyze data collected by “traditional” science in a systematic way, and are considered a part of data science, a new form of science with specific characteristics as discussed as follows [PF13]:

*“At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. Possibly the most closely related concept to data science is data mining — the actual extraction of knowledge from data via technologies that incorporate these principles.”* (Provost and Fawcett, 2013, p.2)

Within the context of this new data science, human or computer errors during data entry or transmission are still regarded as errors, and missing data and sampling distortion are also regarded as errors [HK06] [Han07]. The discrepancies between the results from running a data mining and classification algorithm on a dataset and the actual result from observation or measurement of the dataset are also regarded as errors, classification errors, which include training errors and testing errors, which are a part of prediction errors in statistical terms [HTF01].

While a detailed discussion on errors in semantic and statistical terms is beyond the scope of this research, classification errors have been studied in a qualitative and quantitative way, to serve as a driving force to seek further understanding and knowledge, which leads to this chapter’s topic on ambiguous values and error-sensitive attributes, and one primary question to investigate is – Are there specific attributes in a dataset which may be more vulnerable to cause ambiguity and errors, and if yes, can they be identified in a systematic way?

The rest of this chapter is organized as follows. Section 3.1 introduces three terminologies that are used specifically in this evaluation process of error-sensitive attributes. Section 3.2 explains some initial thoughts and assumptions that lead to this evaluation idea.

Section 3.3 provides detailed information about the proposed evaluation. Section 3.4 summarizes the experiments on five datasets. Section 3.5 discusses the experiment results and the progress of related ideas. Section 3.6 reviews some early influential research work that helped establish the basis of this evaluation proposal, and Section 3.7 concludes the current evaluation development.

### **3.1 Key Concepts**

To introduce this research work with a simplistic and hypothetical scenario, data classification was carried out for a dataset of 10,000 sample instances and 50 data attributes, and the result showed 100 misclassification errors. Our research question based on this hypothetical scenario is, which attributes may be considered the most error-sensitive attributes of the 50 attributes? In order to expand the scope of this specific question to other datasets in general, the above question can be rephrased as – how to identify error-sensitive attributes of a dataset in an effective way? One possible benefit from identifying such error-sensitive attributes can be to help develop error reduction measures in a more specific and attribute-oriented way. Another possible benefit is to raise stakeholders’ awareness of these specific attributes, to further investigate possible special relationships between attributes and errors in a more effective way.

As an attempt to address this question, this study explores and identifies the potentially error-sensitive attributes using three specific terms, ambiguous value range, attribute error counter and error-sensitive attribute. These three terms are used frequently in this research discussion and they all refer to binary classification scenarios within the context of this study, and each attribute of a dataset can potentially be associated with these three



terms. While some of the ideas and experiments have been summarized in an initial report [WZ13], more analytic details are discussed in the next few sections.

For a specific attribute in a dataset, the term *ambiguous value range* describes a value range in which attribute values of negative and positive samples co-exist. It is a grey area of values in which one value range of negative samples overlaps one value range of positive samples, therefore ambiguity and uncertainty arise because this value range cannot determine the class label of occupying samples under this specific attribute.

The term *attribute-error counter* refers to a specific attribute in a dataset with a connection to the previous term *ambiguous value range* and describes the number of misclassified samples with their attribute values being within an ambiguous value range of this specific attribute.

The term *error-sensitive attribute* describes an attribute that is considered to be more prone to errors when compared to others during a data mining and classification process. Its risk assessment is initially based on counting its *attribute-error counter* value, as previously described, and also taking this count on other attributes, then comparing, ranking and selecting the attribute with the highest count as the most error-sensitive attribute.

It is recognized that there is still room for improvement when defining these three new concepts in terms of scope and precision, but when they are used in the context of this study, they can be considered relevant, practical and adequate within a limited scope. Further details about these terms are discussed in the next few sections.

### 3.2 Initial Assumptions

To start with, only attributes with numeric values are considered, and it is expected that there will be no missing values. Such a narrow scope will simplify the formulation of value ranges and the calculation of the ranking between attributes, and will maintain a level of consistency and continuity amongst the different datasets throughout this discussion.

Continuing the brief scenario described in Section 3.1, for the 9,900 samples that have been correctly classified, let us assume there are two simplistic situations for the value ranges of the negative class and positive class within each data attribute to commence this discussion:

1. The first situation is an overlap situation in which there is a grey area of ambiguity between the attribute values of some negative and some positive instances, as shown by the blue overlap area between the **green** oval-shaped negative value area and the **red circle** positive value area on the left side of Figure 3.1;
2. The second situation is a clear-cut separation between the attribute value range of negative instances and the value range of positive instances, as shown by the **green oval-shaped** area and the **red circle** area on the right side of Figure 3.1

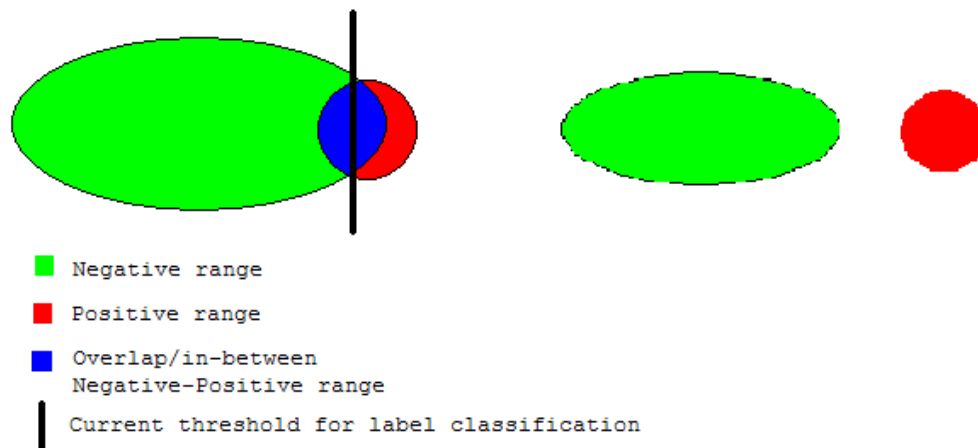


Figure 3.1 – Two attribute value range situations

Based on these two situations in Figure 3.1, the values of an attribute can be grouped into two possible types in relation to binary classification errors:

1. Ambiguous value type - attribute values are inside an ambiguous overlap range – the **blue** area, where some negative instances as well as some positive instances have their attribute values sitting in this range;
2. Unambiguous value type - negative instances have a different attribute value range from positive instances, e.g. negative in the **green** area and positive in the **red** area, so the two class labels are clearly separated and there is no overlap between them.

For a classification error associated with an ambiguous value type in an attribute, it is either a false negative or a false positive. While such an error may not necessarily be caused solely because its attribute value is inside an ambiguous overlapping range, however, such ambiguity may likely increase its risk of being misclassified. Based on this assumption and the earlier definition for the term error-sensitive attribute, if such an attribute whose ambiguous values are associated with a higher count of classification errors, this attribute can be considered more error-sensitive than other attributes which are associated with lower counts of errors.

There are other situations, for example, when the value range of negative instances is almost the same as the value range of positive instances, or, one value range of one class label is fully covered by the value range of the other class label like a total eclipse. These situations may indicate no association between class label distinction and value range separation, which means the specified value range of an attribute in focus may not be closely associated with the errors, or, may only contribute to one of the many reasons that are

associated with errors during a classification process. Other possible factors may include, for example, the correlation impact when multiple attributes are involved with the errors, the propagation of errors from other attributes or conditions, the overshadowing effect from other data components and the environment, and the possibility of sample data contamination, noisy data, or misreading values, etc. [Tay96] [Yoo89]

### 3.3 How the Evaluation Process Works

While this evaluation process seems to share some similarity with the filter type feature selection process, one main difference is that its specific approach can be considered as a post-classification routine - a new third-phase process with a reference to the well-defined two-phase approach [LMS10], as illustrated in Figure 3.4 of Section 3.6.2 Related Work on page 72, to evaluate the resulting errors from the second phase of a classification process. This classification process may already have incorporated its pre-determined feature selection routine. This new third-phase process explores the attribute-error relationship and error-reduction measures, as illustrated in Figure 3.2:

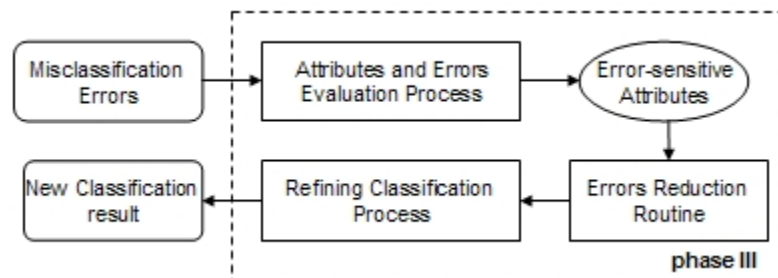


Figure 3.2 - Attribute and Error Evaluation as phase III of the classification process

This evaluation begins as a post-classification analysis. For each attribute used in the data classification, the lower bound and upper bound value of the correctly classified

negative and positive instances are identified, as is the mean value of these two correctly classified groups of samples. After this, each attribute's ambiguous value range can be identified by comparing the lower and upper bound values of each class identified. If an overlapping range is found while comparing these two value ranges, it can subsequently be called an *ambiguous value range* as described in Section 3.1 and 3.2, and ambiguity arises from the fact that both negative and positive instances have attribute values co-existing inside this same value range.

For each attribute containing such an ambiguous value range, a check of the attribute values of all the misclassified sample instances is conducted. If a misclassified instance's attribute value is within this ambiguous value range, then one is added to the attribute-error counter of this attribute. After calculating the attribute-error counter value for all the involved attributes, the attribute with the highest count can be considered as the most error-sensitive attribute. This is because its count of misclassification errors being higher than the other attributes indicates this attribute has a higher level of risk of being misclassified, therefore it is more prone and more sensitive to errors than the others. Alternatively, the ratio of the attribute-error counter to the number of instances with an ambiguous value may also be used to measure such an error-sensitivity level. This latter approach may seem to be following a similar progression path from information gain-to-gain ratio; however, additional work will be required to examine such an alternative theoretically and statistically.

One way to illustrate this evaluation method is through the following brief scenario and process summary.

A supervised classification process is performed on a binary dataset with  $t$  attributes and a total of  $(m + n)$  sample instances. On completion of the first round of the classification process,  $m$  samples are correctly classified and  $n$  misclassification errors are identified, hence the proposed error evaluation process can subsequently begin as a post-classification analysis routine to examine the classification result and errors:

**Input:** A binary dataset of  $(m + n)$  sample instances containing  $t$  attributes with  $m$  samples correctly classified and  $n$  samples misclassified during the initial classification process

**Output:** A ranking list of the  $t$  attributes sorted by their attribute-error counter values from high to low; initial error-reduction measure can be drafted based on the evaluation result and tested by a re-run of the classification process with the new measure

**Evaluation process in three steps:**

Step-1: calculate ambiguous value ranges and attribute-error counter values

for 1 to  $t$  attributes

for 1 to  $m$  correctly classified samples

find negative samples' min & max value and their mean value;

find positive samples' min & max value and their mean value;

compare these min/max values to determine the ambiguous value range;

for 1 to  $n$  misclassified samples

if the attribute value is within its attribute's ambiguous value range

then add one to its attribute's attribute-error counter;

end;

Step-2: sort the attribute-error counter values from high to low to highlight the most error-sensitive attributes in the top-ranking positions;

Step-3: compare this error-sensitive attribute ranking list with other attribute ranking lists, such as the info-gain and gain-ratio ranking list, to evaluate the most error-sensitive attributes and possible error-reduction measures.

One critical issue in this evaluation process is how to determine the lower and upper bound of the ambiguous value range for each attribute. In the current stage of this study, only two simplified overlap situations are discussed here, as shown in the middle and right diagram of Figure 3.3:

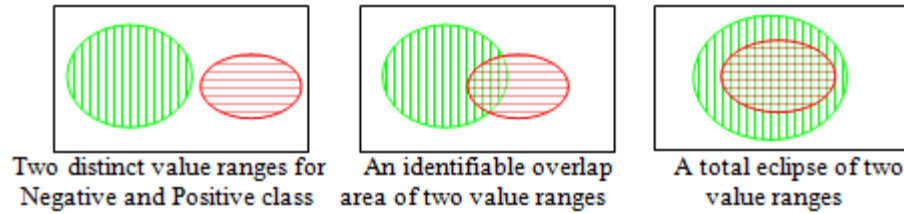


Figure 3.3 – Three value range examples

For the first overlap situation when the overlap area is small and distinguishable as shown in the middle diagram, the ambiguous value range can be determined by identifying the minimum and maximum attribute value of the negative and positive class and using them as the lower and upper boundary. For the second overlap situation when it is less distinguishable, as shown in the right diagram, the ambiguous value range can be determined by using the mean value of the correctly classified negative and the positive samples' attribute values as the lower and upper boundary instead.

One hypothesis which arises from this evaluation idea is, if the most error-sensitive attribute is not the most significant attribute that has been used by the underlying classification model, then the removal or weighting down one or two of the most error-sensitive attributes may help reduce the classification error rate.

To test whether our proposed evaluation process can actually identify the most error-sensitive attributes, experiments were carried out on a number of real-world datasets. These experiments and their analyses are discussed in the next two sections.

### 3.4 Experiments

Initial experiments were performed using the J48 decision tree classifier in WEKA [HFH09] and its standard system configuration, e.g. minimum description length (MDL)

correction method is used for pruning to reduce “overfitting”, confidence factor is 0.25 as defined by default to optimize pruning because confidence factor  $> 0.5$  would actually mean no pruning, minimum number of instances per leaf is 2 as defined by default to be the minimum size of a leaf and also the data separation and branch split threshold, and the test option is 10-fold cross-validation to maximize the number of training samples and testing records within the available records in the datasets.

### 3.4.1 The Pima Indians Diabetes Dataset

There are eight attributes and 768 sample instances in this dataset, 500 of them with their class label pre-determined as 0 meaning it tested negative for diabetes, and 268 with their attribute class label pre-determined as 1 meaning it tested positive for diabetes, and there were no missing attribute values. The initial classification process using the J48/C4.5 classifier in WEKA correctly classified 568 instances. The results are summarized in the following table:

Table 3-1 – Pima dataset’s initial classification result by J48 in WEKA

Class Label	Classified negative	Classified positive	<i>Actual class count</i>
Tested negative	407 (TN)	93 (FP)	<i>500</i>
Tested positive	107 (FN)	161 (TP)	<i>268</i>
Correctly classified instances: 568 ~ 73.96%			<i>768</i>
Misclassification errors: 200 ~ 26.04%			

This means that there are 93 false positive errors, 107 false negative errors, and a total of 200 misclassification errors. Our research question in this experiment can then be amended accordingly to be: “What is the top one or top two most error-sensitive attributes of the eight attributes in the Pima diabetes dataset?”



Using the proposed evaluation method as a part of the post-classification analysis, each attribute's value range of the correctly classified negative (TN) and positive samples (TP) is identified and their ambiguous value range and attribute-error counter are also calculated by comparing their lower and upper bound values, as shown on the left-hand side of Table 3-2. For this first method of attribute-error counter calculation, the ambiguous value range is determined by comparing the minimal and maximal value of the correctly classified positive (TP) samples and the correctly classified (TN) negative samples, and taking the higher value of the two minimums as the lower bound of the ambiguous value range, and the lower value of the two maximums as the upper bound of the ambiguous value range.

Table 3-2 - Two methods of calculation for the Pima dataset's ambiguous value range

<b>First method</b> of measuring attribute-error counters based on the lower and upper bound of TN and TP attribute values				<b>Second method</b> of measuring attribute-error counters based on the mean values of TN and TP		
Attribute	Value range of TN	Value range of TP	Attribute-error counter	Mean value of TN	Mean value of TP	Attribute-error counter (alt)
<i>preg</i>	<b>0~13</b>	<b>0~17</b>	199 ( <b>0~13</b> )	3.30	4.87	<b>20</b> (3.30~4.87)
<i>plas</i>	0~ <b>175</b>	<b>92~199</b>	174 ( <b>92~175</b> )	109.98	141.26	<b>83</b> (110~141)
<i>pres</i>	0~ <b>122</b>	<b>0~114</b>	200 ( <b>0~114</b> )	68.18	70.82	<b>18</b> (68~71)
<i>skin</i>	<b>0~54</b>	0~99	199 ( <b>0~54</b> )	19.66	22.16	<b>7</b> (20~22)
<i>insu</i>	0~ <b>485</b>	<b>0~846</b>	197 ( <b>0~485</b> )	68.79	100.34	<b>15</b> (69~100)
<i>mass</i>	0~ <b>57.3</b>	<b>23.3~59.4</b>	195 ( <b>23.3~57.3</b> )	30.30	35.14	<b>70</b> (30~35)
<i>pedi</i>	0.08~ <b>1.73</b>	<b>0.09~2.42</b>	197 ( <b>0.09~1.73</b> )	0.43	0.55	<b>24</b> (0.43~0.55)
<i>age</i>	21~81	<b>22~70</b>	195 ( <b>22~70</b> )	31.19	37.07	<b>31</b> (31~37)

It is observed that for half of the eight attributes, *preg*, *pres*, *skin* and *age*, the value overlap ranges between their correctly classified negative and positive samples, reflecting the *total eclipse* situation illustrated on the right-hand side of Figure 3.3. As for the other four attributes, *plas*, *insu*, *mass* and *pedi*, the value overlap ranges between their correctly

classified negative and positive samples, reflecting the *identifiable overlap* situation illustrated in the middle diagram in Figure 3.3.

The outcome from the first method of attribute-error counter calculation is unclear. Seven attribute-error counters reached almost 200, with 200 being the total number of all misclassification errors. The only exception is the attribute *plas*, but its counter is still valued at 174, which is quite close to 200, as shown in the fourth column, Attribute-error counter in Table 3-2. This means, using the combination of the lowest and highest attribute values between the correctly classified negative and positive samples as the ambiguous value range boundaries does not help identify the most error-sensitive attributes in the Pima diabetes dataset.

An alternative method is to use the mean value of the negative and positive samples in each attribute as the lower and upper boundary of its respective ambiguous value range, as shown on the right-hand side of Table 3-2. The attribute-error counters can then be calculated based on the revised ambiguous value ranges. As shown in the last column in Table 3-2, four distinctive groups of attributes are then easily identified, with the two highest counts being 83 and 70 for the attributes *plas* and *mass*, and the two lowest counts being 7 and 15 for the attributes *skin* and *insu*.

Other alternative methods can also be considered, for example, using the median or mode value instead of the mean value, or a proportional average value based on stratified sampling in the negative and the positive samples. These alternatives are discussed and evaluated in the next stage of the study and development.

Once the ambiguous value ranges and attribute-error counters are calculated, a list of attributes ranked by their attribute-error counters can be produced, together with the other ranking lists of attributes, for example, one by the gain ratio method [Qui86] and one by the information gain method [Sha48]. When comparing this attribute-error count ranking list and other attribute selection and ranking lists, the variation between the most error-sensitive attributes identified by ranking the attribute-error counters and the other most significant attributes identified by the other algorithms such as gain ratio and information gain can also be analysed to help gain an increased understanding of the impact of ambiguous value ranges and errors and their potential association with other determining factors such as information gain and gain ratio.

As shown in Table 3-3, the three most error-sensitive attributes, *plas*, *mass* and *age* identified by ranking the attribute-error counters, are also the top three attributes ranked by the gain ratio and information gain algorithm implemented in WEKA. As a way to provide an overall contrast based on attribute value variation and distribution, the ratio value between the number of different/distinct values of this attribute and the number/percentage of instances that have an attribute value that is unique is also listed in the last column named “Distinct values / Unique samples (WEKA)” as background information. More details on the distinct and unique function in WEKA is discussed in Section 3.5.4.

Having compared these three ranking lists to gather more information, the next step is to examine if the classification accuracy can be enhanced by developing some error-reduction features after identifying the most error-sensitive attributes. One way to test this

idea is to filter out the most error-sensitive attributes and re-run the classification with the dataset.

Table 3-3 – The Pima dataset’s attribute-error counter, gain ratio and information gain ranking lists

Rank	Attribute-error counter	Gain ratio	Info gain	Distinct values / Unique samples (WEKA)
1	plas (83)	plas (0.0986)	plas (0.1901)	pedi (517 / 346 ... 45%)
2	mass (70)	mass (0.0863)	mass (0.0749)	mass (248 / 76 ... 10%)
3	age (31)	age (0.0726)	age (0.0725)	insu (186 / 93 ... 12%)
4	pedi (24)	preg (0.0515)	insu (0.0595)	plas (136 / 19 ... 2%)
5	preg (20)	insu (0.0394)	skin (0.0443)	age (52 / 5 ... 1%)
6	pres (18)	pedi (0.0226)	preg (0.0392)	skin (51 / 5 ... 1%)
7	insu (15)	skin (0.0224)	pedi (0.0208)	pres (47 / 8 ... 1%)
8	skin (7)	pres (0.0144)	pres (0.014)	preg (17 / 2 ... 0%)

In this experiment with the Pima diabetes dataset, the attributes *plas* and *mass* are identified as the two most error-sensitive attributes by the proposed evaluation process because their attribute-error counter values are more than double other attributes, and subsequently, these two attributes are removed from the classification process as an error-reduction measure. A comparison between the results of the initial J48/C4.5 classification run and a re-run of the classification without the two most error-sensitive attributes is given in the form of a confusion matrix comparison in Table 3-4.

While this re-run result does not seem supportive at first glance, a good reason is identified for this worse-than-expected result, as explained in the discussion section.

Table 3-4 - Comparing Pre- vs. Post- removal of error-sensitive attributes in the Pima dataset

Original C4.5 classification with eight attributes	Re-run C4.5 classification without the two most error-sensitive attributes <i>plas</i> and <i>mass</i>
=== Stratified cross-validation === Correctly classified samples    568 <b>73.96%</b> Incorrectly classified samples   200 26.04% === Confusion Matrix === a   b  <-- classified as 407 93    a = tested_negative 107 161    b = tested_positive	=== Stratified cross-validation === Correctly classified samples    515 <b>67.06%</b> Incorrectly classified samples   253 32.94 % === Confusion Matrix === a   b  <-- classified as 404 96    a = tested_negative 157 111    b = tested_positive

### 3.4.2 The Wisconsin Breast Cancer Dataset

In the Wisconsin breast cancer dataset, there are nine meaningful attributes and 699 samples, and 458 samples have been pre-determined as benign in the attribute *class* meaning negative, and 241 samples as malignant meaning positive. It is noted that 16 samples are missing their *Bare Nuclei* attribute values, but because only a small number of samples are impacted, the initial classification process using the J48 classifier in WEKA with 10-fold cross-validation proceeds as normal, and the result is summarized in Table 3-5.

Table 3-5 - Wisconsin dataset's initial classification result

Class label	Classified benign/negative	Classified malignant/positive	<i>Actual class count</i>
Negative	434 (TN)	24 (FP)	<b>458</b>
Positive	17 (FN)	224 (TP)	<b>241</b>
Correctly classified instances: 658 ~ 94.13%			<b>699</b>
Misclassification errors: 41 ~ 5.87%			

This means that there are 41 classification errors, 24 of them being false positive and 17 being false negative. One specific question related to this study about this classification process is, “What is the top one or top two most error-sensitive attributes of the nine meaningful attributes in this Wisconsin cancer dataset?”

Two calculation methods to produce the ambiguous value ranges and attribute-error counters, as described in the earlier section for the Pima diabetes dataset experiment, are applied to the classification result as a part of the post-classification analysis, with the evaluation and comparison result outlined in Table 3-6.

Table 3-6 - Two methods of calculation for the Wisconsin dataset's ambiguous value range

<b>First method</b> of measuring attribute-error counters based on the lower and upper bound of TN and TP attribute values				<b>Second method</b> (alt) of measure based on the mean values of TN and TP attribute values		
Attribute	Value range of TN	Value range of TP	Attribute-error counter	Mean value of TN	Mean value of TP	Attribute-error counter (alt)
Clump Thickness	1~8	1~10	<b>38</b> (1~8)	2.96	7.20	<b>27</b> (2.96~7.20)
Uniformity of Cell Size	1~4	1~10	<b>34</b> (1~4)	1.33	6.57	<b>32</b> (1.33~6.57)
Uniformity of Cell Shape	1~4	1~10	<b>29</b> (1~4)	1.44	6.56	<b>32</b> (1.44~6.56)
Marginal Adhesion	1~6	1~10	<b>35</b> (1~6)	1.36	5.58	<b>21</b> (1.36~5.58)
Single Epithelial Cell Size	1~10	2~10	<b>38</b> (2~10)	2.12	5.30	<b>20</b> (2.12~5.30)
Bare Nuclei	1~10	1~10	<b>41</b> (1~10)	1.35	7.63	<b>21</b> (1.35~7.63)
Bland Chromatin	1~7	1~10	<b>40</b> (1~7)	2.10	5.98	<b>25</b> (2.10~5.98)
Normal Nucleoli	1~8	1~10	<b>37</b> (1~8)	1.29	5.86	<b>15</b> (1.29~5.86)
Mitoses	1~8	1~10	<b>40</b> (1~8)	1.06	2.99	<b>3</b> (1.06~2.99)

Using the first calculation method based on the lower and upper bound of the true negative and true positive attribute values, seven attribute-error counters are valued between 35 and 41, almost same as the total number of errors of 41. As for the other two attribute-error counters, one is valued at 29 which is almost 30, and the other one is valued at 34, as shown in the fourth column of Table 3-6. This is another indication that using the minimal and maximal attribute values of the true negative and true positive samples as the lower and upper bound of an attribute's ambiguous value range might not be the most suitable measurement method for some datasets.

In comparison, the second method based on the mean values of true negative and true positive attribute values is able to produce distinct attribute-error counters to differentiate the nine meaningful attributes, making it clear that the attributes *Uniformity of Cell Size* and *Uniformity of Cell Shape* are the two most error-sensitive attributes, as shown in the last column of Table 3-6.

Further work has been undertaken to generate the gain ratio and information gain ranking list for comparison with the error-sensitive ranking list, as shown in Table 3-7. This comparison indicates that *Uniformity of Cell Size* and *Uniformity of Cell Shape* are identified as the two most error-sensitive attributes by the proposed evaluation method, and encouragingly, the information gain algorithm also identifies these two attributes as the two most significant attributes. However, the gain ratio algorithm identifies two different attributes, *Normal Nucleoli* and *Single Epithelial Cell Size*, as the two most significant attributes.

This leads to one question, why are the two most error-sensitive attributes ranked by the attribute-error counter in the Pima diabetes dataset also the two most significant attributes ranked by the information gain algorithm and the gain ratio algorithm, but the gain ratio algorithm returns two different attributes in the Wisconsin dataset even though the information gain algorithm and the attribute-error counter ranking method return the same two attributes? The exploration of this question is analyzed in the experiment discussion section of this chapter.

Table 3-7 - Wisconsin dataset's attribute-error counter, gain ratio and information gain ranking lists

Rank	Attribute-error counter	Gain ratio	Info gain	Distinct values/Unique samples (WEKA)
1	Uniformity of Cell Size (32)	Normal Nucleoli (0.399)	Uniformity of Cell Size (0.675)	Uniform of Cell Size (10/0...0%)
2	Uniformity of Cell Shape (32)	Single Epithelial Cell Size (0.395)	Uniformity of Cell Shape (0.66)	Uniform of Cell Shape (10/0...0%)
3	Clump Thickness (27)	Uniformity of Cell Size (0.386)	Bare Nuclei (0.564)	Bare Nuclei (10/0...0%)
4	Bland Chromatin (25)	Bare Nuclei (0.374)	Bland Chromatin (0.543)	Bland Chromatin (10/0...0%)
5	Marginal Adhesion (21)	Uniformity of Cell Shape (0.314)	Single Epithelial Cell Size (0.505)	Sing Epithelial Cell Size (10/0...0%)
6	Bare Nuclei (21)	Bland Chromatin (0.303)	Normal Nucleoli (0.466)	Normal Nucleoli (10/0...0%)
7	Single Epithelial Cell Size (20)	Mitoses (0.299)	Clump Thickness (0.459)	Clump Thickness (10/0...0%)
8	Normal Nucleoli (15)	Marginal Adhesion (0.271)	Marginal Adhesion (0.443)	Marginal Adhesion (10/0...0%)
9	Mitoses (3)	Clump Thickness (0.21)	Mitoses (0.198)	Mitoses (9/0...0%)

In order to verify if the most error-sensitive attributes identified by the attribute-error counter evaluation process are indeed key contributors to the misclassification errors, and if the filtering of these most error-sensitive attributes is an effective error-reduction measure in a classification process, a re-run of the J48/C4.5 decision tree classification without these two attributes is conducted, and a comparison between the initial J48/C4.5 classification run and a re-run of the classification without the two most error-sensitive attributes is given in the form of a confusion matrix comparison in Table 3-8.



Table 3-8 - Pre- vs. Post- removal of error-sensitive attributes in the Wisconsin dataset

Original C4.5 classification with nine attributes	Re-run C4.5 classification without the two most error-sensitive attributes
==== Stratified cross-validation ====	==== Stratified cross-validation ====
Correctly Classified Instances 658 94.13%	Correctly Classified Instances 669 95.71%
Incorrectly Classified Instances 41 5.87%	Incorrectly Classified Instances 30 4.29%
==== Confusion Matrix ====	==== Confusion Matrix ====
a b <-- classified as	a b <-- classified as
434 24   a = benign	440 18   a = benign
17 224   b = malignant	12 229   b = malignant

This test result shows that classification accuracy increased from 94.13% in the initial run with 9 attributes to 95.71% in the re-run without the two most error-sensitive attributes, which is almost a 2% accuracy improvement and indicates 11 more patients may subsequently be correctly diagnosed. This improvement appears to confirm that the removal of the two error-sensitive attributes can be somewhat an effective error reduction measure in this Wisconsin dataset based on the evaluation result, so this can be considered as a supportive test case for the proposed evaluation and its identification of the two most error-sensitive attributes.

The Wisconsin dataset and the Pima diabetes dataset are two UCI binary datasets that can be tested directly with this proposed evaluation process without the need to modify their data values. Three other UCI multi-class datasets have been modified slightly to suit the requirements of binary classification and have subsequently been used in three follow-up experiments to verify the issues discovered in the two earlier test cases.

### 3.4.3 The Ecoli Dataset

There are 336 samples, seven meaningful attributes and up to eight class labels in this dataset. The eight class labels have subsequently been re-grouped and consolidated into two class labels, *membrane* as negative, and *non\_membrane* as positive, in accordance with the class distribution described by the original dataset owners.

After performing an initial round of classification using the J48/C4.5 decision tree classifier in WEKA and applying the proposed evaluation as a part of the post-classification analysis, a ranking list of error-sensitive attributes is generated based on the mean value of the correctly classified negative and positive samples, and this list is used to compare with the gain ratio and information gain attribute ranking lists to help explore the possible interrelation between the most error-sensitive attributes and the most significant attributes, as shown Table 3-9:

Table 3-9 - Ecoli dataset's attribute-error counter, gain ratio and information gain ranking lists

Rank	Attribute-error counter	Gain ratio	Info gain
1	chg (17)	alm1 (0.3999)	alm1 (0.627)
2	alm1 (15)	alm2 (0.3683)	alm2 (0.563)
3	lip (14)	lip (0.1977)	aac (0.26)
4	aac (8)	aac (0.1844)	mcb (0.173)
5	alm2 (5)	mcb (0.1312)	gvh (0.053)
6	gvh (3)	gvh (0.0915)	lip (0.038)
7	mcb (1)	chg (0)	chg (0)

One interesting point in this Ecoli dataset experiment is that the attribute *chg* is evaluated as the most error-sensitive attribute, but this same attribute is ranked the least significant by the gain ratio algorithm and the information gain algorithm. Meanwhile,

*alm1* is ranked as the second most error-sensitive attribute and is also ranked as the most significant attribute by the gain ratio algorithm as well as the information gain algorithm. Because of this interesting ambiguity, two possible error-reduction measures are tested, one is to filter out the *chg* attribute and the other is to filter out the top three most error-sensitive attributes, *chg*, *alm1* and *lip*, and their result comparison is given in the form of confusion matrix comparison in Table 3-10:

Table 3-10 - Pre- vs. Post- removal of error-sensitive attributes in the Ecoli dataset

Original C4.5 classification with all seven attributes	Re-run C4.5 classification without the most error-sensitive at- tribute, <i>chg</i>	Re-run C4.5 classification without the three most error-sensitive attrib- utes, <i>chg</i> , <i>alm1</i> and <i>lip</i>
=== Stratified cross-validation === Correctly Classified: 318 94.64% Incorrectly Classified: 18 5.36%	=== Stratified cross-validation === Correctly Classified 318 94.64% Incorrectly Classified: 18 5.36%	=== Stratified cross-validation === Correctly Classified 311 92.56% Incorrectly Classified: 25 7.44%
=== Confusion Matrix === a b <-- classified as 189 6   a = membrane 12 129   b = non_membrane	=== Confusion Matrix === a b <-- classified as 189 6   a = membrane 12 129   b = non_membrane	=== Confusion Matrix === a b <-- classified as 191 4   a = membrane 21 120   b = non_membrane

This comparison indicates that the removal of the most error-sensitive attribute *chg* does not have any impact on the re-run of the classification for this Ecoli dataset and the removal of the top three error-sensitive attributes, *chg*, *alm1* and *lip*, have actually decreased the accuracy rate from 94.64% to 92.56%. This result shares some similarity with the outcome from the Pima diabetes dataset experiment but differs from the Wisconsin dataset. The possible causes are discussed in Section 3.5.

### 3.4.4 The Liver Disorders Dataset

There are 345 samples and seven attributes in this dataset and the attribute *selector* can be regarded as a class label according to the dataset owners' description, and its binary

*selector* value of 1 can be relabeled as negative, and other binary value of 2 can be relabeled as positive. After performing an initial round of classification using the J48/C4.5 decision tree classifier in WEKA and applying the proposed evaluation as a part of the post-classification analysis, a ranking list for the error-sensitive attributes in this liver disorders dataset is generated and compared with the gain ratio ranking list and information gain ranking list, as shown in Table 3-11.

The reason why the gain ratio and information gain ranking value format looks different in this Table 3-11 is, when testing the attribute selection methods, WEKA's default sampling strategy is to use a full training set, which works well for most datasets except for the Liver Disorder dataset because most of its attribute ranking values show a zero value under this default sampling strategy, therefore a cross-validation sampling strategy is used which leads to this change of ranking value format.

Table 3-11 - Liver Disorders dataset's attribute-error counter, gain ratio and information gain ranking lists

Rank	Attribute-error counter	Gain ratio	Info gain
1	sgot (22)	gammagt (0.097 +- 0.065)	gammagt (0.051 +- 0.013)
2	mcv (16)	alkphos (0 +- 0)	alkphos (0 +- 0)
3	alkphos (10)	mcv (0 +- 0)	mcv (0 +- 0)
4	sgpt (8)	drinks (0 +- 0)	drinks (0 +- 0)
5	gammagt (2)	sgpt (0.006 +- 0.018)	sgpt (0.003 +- 0.01)
6	drinks (0)	sgot (0.013 +- 0.04)	sgot (0.003 +- 0.009)

The comparison of these three ranking lists shows that the attribute *sgot* is the most error-sensitive attribute as ranked by the attribute-error counter, but it is also the least significant attribute ranked by the gain ratio algorithm as well as the info gain algorithm. On the other hand, the attributes *mcv* and *alkphos* are ranked as the second and the third most

error-sensitive attributes, and they are also ranked the second and the third most significant attributes by the gain ratio and information gain methods. Because of this interesting contrast between these three attributes, the J48/C4.5 decision tree classification is re-run numerous times using different filtering combinations from these three attributes to test for possible error-reduction measures.

Two tests with enhanced classification results are selected for presentation to highlight some potential benefits in identifying error-sensitive attributes and developing alternate error-reduction measures by the proposed evaluation process. In one test, the attribute *sgot* is filtered, in another test, both *sgot* and *mcv* are filtered, and their result comparison with the initial classification is outlined in the form of a confusion matrix, as shown in Table 3-12.

Table 3-12 - Pre- vs. Post- removal of error-sensitive attributes in the Liver Disorders dataset

Original C4.5 classification with all seven attributes	Re-run C4.5 classification without the most error-sensitive attribute, <i>sgot</i>	Re-run C4.5 classification without the top two most error-sensitive attributes, <i>sgot</i> and <i>mcv</i>
=== Stratified cross-validation === Correctly Classified: 233 67.54% Incorrectly Classified: 112 32.46% === Confusion Matrix === a b <-- classified as 72 73   a = 1 39 161   b = 2	=== Stratified cross-validation === Correctly Classified: 237 68.70% Incorrectly Classified: 108 31.30% === Confusion Matrix === a b <-- classified as 82 63   a = 1 45 155   b = 2	=== Stratified cross-validation === Correctly Classified: 243 70.43% Incorrectly Classified: 102 29.57% === Confusion Matrix === a b <-- classified as 78 67   a = 1 35 165   b = 2

This comparison indicates that in one test, some marginal improvement is gained from the initial accuracy rate of 67.54% to 68.70% after the single most error-sensitive attribute *sgot* is filtered out from the classification process. In another test, a more significant improvement is gained when its accuracy rate is lifted to 70.43% after the top two

most error-sensitive attributes, *sgot* and *mcv*, are filtered out from the classification process. While these two test cases show supportive results for the proposed evaluation, it is also true that the results obtained using other filtering combinations of these three attributes are not as appealing or supportive.

### 3.4.5 The Page Blocks Dataset

This is another interesting realistic UCI dataset for testing with ten attributes, 5,473 samples and five classes, *text*, *horizontal line*, *vertical line*, *picture* and *graphic*, which are consolidated into two class labels, *text* as positive and *non\_text* as negative, to become part of the binary classification scenarios in this study.

Table 3-13 - Page Blocks dataset's attribute-error counter, gain ratio and information gain ranking lists

Rank	Attribute-error counter	Gain ratio	Info gain
1	mean_tr (89)	length (0.2993)	height (0.2357)
2	p_black (43)	height (0.1494)	mean_tr (0.1828)
3	eccen (29)	mean_tr (0.143)	wb_trans (0.1537)
4	height (25)	wb_trans (0.1129)	p_black (0.1518)
5	blackpix (15)	p_black (0.1082)	eccen (0.1447)
6	area (10)	area (0.0942)	p_and (0.1364)
7	length (9)	p_and (0.0775)	blackand (0.0899)
8	blackand (6)	eccen (0.0752)	area (0.0893)
9	p_and (2)	blackpix (0.0513)	length (0.0605)
10	wb_trans (0)	blackand (0.0438)	blackpix (0.0549)

After an initial classification process, a post-classification analysis is conducted which includes the proposed error-sensitive attribute evaluation process. An attribute-error counter ranking list is generated by this evaluation, together with the generation of an information gain ranking list and a gain ratio ranking list using WEKA for comparison

purposes to help explore and understand value pattern similarity and the discrepancies between the most error-sensitive attributes and the most significant attributes, as shown in Table 3-13.

This comparison shows that attribute *mean\_tr* is the most error-sensitive attribute and it is also ranked as the second and the third most significant attribute by the info gain algorithm and gain ratio algorithm, respectively. Attributes *p\_black* and *eccen* are the second and third most error-sensitive attributes and they are ranked the fifth and eighth significant attributes by the gain ratio algorithm, and are ranked the fourth and fifth by the information gain algorithm. At a glance, these three ranking lists all look rather different from each other, but the attribute-error counter ranking list is the most at odds with the other two.

Table 3-14 - Pre- vs. Post- removal of error-sensitive attributes in Page Blocks dataset

Original C4.5 classification with all ten attributes	Re-run C4.5 classification without the most error-sensitive at- tribute, <i>mean_tr</i>	Re-run C4.5 classification without the second and third most error-sen- sitive attributes, <i>p_black</i> and <i>eccen</i>
==== Stratified cross-validation ==== Correctly Classified: 5321 97.22% Incorrectly Classified: 152 2.78% ==== Confusion Matrix ==== a b <-- classified as 4844 69   a = text 83 477   b = non_text	==== Stratified cross-validation ==== Correctly Classified: 5320 97.20% Incorrectly Classified: 153 2.78% ==== Confusion Matrix ==== a b <-- classified as 4843 70   a = text 83 477   b = non_text	==== Stratified cross-validation ==== Correctly Classified: 5323 97.26% Incorrectly Classified: 150 2.74% ==== Confusion Matrix ==== a b <-- classified as 4844 69   a = text 81 479   b = non_text

Various error-sensitive attribute filtering combinations are tested to see if the error rate can be reduced after the identification of the most error-sensitive attributes, but no major improvement is observed. Two of these tests are selected for presentation in this report, one test has the most error-sensitive attribute *mean\_tr* filtered out, the other test has

the second and third most error-sensitive attributes, *p\_black* and *eccen*, both filtered out. The result comparison is given in the form of a confusion matrix in Table 3-14.

This comparison indicates that the removal of the single most error-sensitive attribute *mean\_tr* in the Page Blocks dataset has little impact on the classification process, as the accuracy rate decreases only a little from 97.22% to 97.20% with one extra false negative error out of the 5,473 sample instances. On the other hand, the removal of the second and third most error-sensitive attributes improves the accuracy rate marginally from 97.22% to 97.26%, showing a reduction of two false positive errors. Further analysis on the result is in the next section.

To summarize the experiments on these datasets, here is an overview of the results with some specific contexts. The Wisconsin dataset and Liver Disorders dataset show supportive results at various levels, and in particular, one test scenario on the Liver Disorders dataset shows that the error rate is reduced by almost 3%. While the Pima diabetes dataset shows some unsupportive results, the worse result being that the error rate actually increased by almost 7%, there are good reasons for these unsupportive results which may potentially turn this into a marginally supportive scenario. The Page Blocks dataset and Ecoli dataset produced neutral results on the surface; but further analysis of the experiments shows that these neutral results may be considered as supportive evidence when taking into account the intricate value patterns of some key attributes and the correlation between them.



### 3.5 Experiment Analysis and Discussion

One key objective of this study is to understand how classification errors may be related to error-sensitive attributes, and how an attribute's ambiguous value range may influence the risk of error in a class label determination process, and an error-sensitive evaluation process has been proposed to explore and assist with such understanding. Of the experiments that have been conducted so far, some show supportive results. Even though the number of successful and supportive test results is limited and in the minority, it is still an encouraging sign to progress this study on value ambiguity and error sensitivity to another level in the future, and a detailed discussion about the results is given in the next few sections.

#### 3.5.1 Evaluating Ambiguous Value Range by Overlapping Minimum-Maximum

The first proposed estimation method for the ambiguous value range is to use the overlapping minimal and maximal value of the negative and positive samples as the value range boundaries, but the experiments demonstrate that this minimum-maximum evaluation method cannot reliably establish an ambiguous value range between a distinctive value range for the negative samples and a distinctive value range for the positive samples. It works for a small number of attributes in a simple and effective way but cannot repeat such success for a larger number of attributes, two such examples being the Pima diabetes dataset and the Wisconsin cancer dataset.

In the Pima diabetes dataset example using the minimum-maximum evaluation method, the ambiguous value range of attribute *plas* accounts for 42% of the overall value range but accounts for 87% of the misclassification errors; the ambiguous value range of attribute *skin* accounts for 55% of the overall value range but accounts for almost 100% of

the misclassification errors; and the ambiguous value range of attribute *mass* accounts for 57% of the overall value range but accounts for 98% of the misclassification errors, as shown in Table 3-15. These higher error counts over smaller value ranges in the ratio of around 2:1 are evidence that the identified ambiguous value ranges are more prone to errors for those three attributes and are almost twice as likely to be associated with misclassification errors.

Table 3-15 shows that the ambiguous value range of attribute *preg* accounts for 76% of the overall value range and accounts for almost 100% of the misclassification errors; and the ambiguous value range of attribute *age* accounts for 80% of the overall value range and accounts for 98% of the misclassification errors. The high percentage of ambiguous value ranges at 75+% over their full value ranges for these two attributes is an indication that the minimum-maximum evaluation method is not an effective way to identify their ambiguous value ranges.

Table 3-15 - Pima diabetes dataset's ambiguous value range by overlapping minimum and maximum of TN & TP attribute values

First method of measuring ambiguous value range based on overlapping minimum & maximum of TN & TP attribute values				
Attribute	Value range of TN	Value range of TP	Attribute-error counter	Value range ratio : Error ratio
preg	<b>0~13</b>	<b>0~17</b>	199 ( <b>0~13</b> )	13/17 = 76% : 199/200 = 100%
plas	0~ <b>175</b>	<b>92~199</b>	174 ( <b>92~175</b> )	83/199 = <b>42%</b> : 174/200 = <b>87%</b>
pres	0~122	<b>0~114</b>	200 ( <b>0~114</b> )	114/122 = 93% : 200/200 = 100%
skin	<b>0~54</b>	0~99	199 ( <b>0~54</b> )	54/99 = <b>55%</b> : 199/200 = <b>100%</b>
insu	0~ <b>485</b>	<b>0~846</b>	197 ( <b>0~485</b> )	485/846 = <b>57%</b> : 197/200 = <b>99%</b>
mass	0~ <b>57.3</b>	<b>23.3~59.4</b>	195 ( <b>23.3~57.3</b> )	34/59.4 = <b>57%</b> : 195/200 = <b>98%</b>
pedi	0.08~ <b>1.73</b>	<b>0.09~2.42</b>	197 ( <b>0.09~1.73</b> )	1.64/2.34 = 70% : 197/200 = 99%
age	21~81	<b>22~70</b>	195 ( <b>22~70</b> )	48/60 = 80% : 195/200 = 98%

In the Wisconsin cancer dataset using the minimum-maximum evaluation method, the ambiguous value range of attribute *Uniformity of Cell Size* accounts for 40% of the overall value range but accounts for 82% of the misclassification errors, and the ambiguous value range of attribute *Uniformity of Cell Shape* accounts for 40% of the overall value range but accounts for 71% of the misclassification errors. While these two attributes show some supportive evidence, other attributes are not so supportive. For example, the ambiguous value range of attribute *Bare Nuclei* accounts for 100% of the overall value range and so, of course, accounts for 100% of the misclassification errors; and the ambiguous value range of attribute *Single Epithelial Cell Size* accounts for 90% of the overall value range and accounts for 93% of the misclassification errors, as shown in Table 3-16.

Table 3-16 - Wisconsin cancer dataset's ambiguous value range by overlapping minimum and maximum of TN & TP attribute values

First method of measuring ambiguous value range based on overlapping minimum & maximum of TN & TP attribute values				
Attribute	Value range of TN	Value range of TP	Attribute-error counter	Value Range ratio : Error ratio
Clump Thickness	1~8	1~10	<b>42</b> (1~8)	8/10 = 80% : 42/45 = 93%
Uniformity of Cell Size	1~4	1~10	<b>37</b> (1~4)	4/10 = <b>40%</b> : 37/45 = <b>82%</b>
Uniformity of Cell Shape	1~4	1~10	<b>32</b> (1~4)	4/10 = <b>40%</b> : 32/45 = <b>71%</b>
Marginal Adhesion	1~6	1~10	<b>39</b> (1~6)	6/10 = 60% : 39/45 = 87%
Single Epithelial Cell Size	1~10	2~10	<b>42</b> (2~10)	9/10 = 90% : 42/45 = 93%
Bare Nuclei	1~10	1~10	<b>45</b> (1~10)	10/10 = 100% : 45/45 = 100%
Bland Chromatin	1~7	1~10	<b>44</b> (1~7)	7/10 = 70% : 44/45 = 98%
Normal Nucleoli	1~8	1~10	<b>41</b> (1~8)	8/10 = 80% : 41/45 = 91%
Mitoses	1~8	1~10	<b>44</b> (1~8)	8/10 = 80% : 44/45 = 98%

One obvious issue with this method of calculating the ambiguous value range using minimum-maximum evaluation is that it does not exclude the outlier values in either the

negative or positive samples when using their minimal and maximal attribute values to establish the ambiguous value ranges for the evaluation. For example, attribute *plas* in the Pima diabetes dataset has only two samples with an attribute value of 0 in the positive class, and the rest of this class has a starting value of 78 onwards. There are three samples with a *plas* attribute value of 0 in the negative class, and the rest of this class has a starting value of 44 onwards. Because both negative and positive samples have a value of 0 for this attribute, 0 is the lower boundary of its ambiguous value range, even though this value of 0 appears to be an outlier value.

A similar example of outlier values can be observed in the attribute *Single Epithelial Cell Size* of the Wisconsin cancer dataset, as shown in Table 3-17. At the **lower end** of the boundary of value 1, there is only one malignant sample but 46 benign samples, hence this value of 1 remains in the ambiguous value range because one hosting sample is a correctly classified positive sample. The upper end of the boundary of value 10 is a similar example with only one correctly classified negative sample but 30 positive samples.

Table 3-17 - Outlier values amongst negative/positive samples in the Wisconsin dataset

Single Epithelial Cell Size value ranging from 1 to 10	Pre-determined class	Instances count
1	benign/negative	46
<b>1</b>	<b>malignant/positive</b>	<b>1</b>
2	benign/negative	363
2	malignant/positive	23
...	...	...
<b>8</b>	<b>benign/negative</b>	<b>2</b>
8	malignant/positive	19
<b>9</b>	<b>malignant/positive</b>	<b>2</b>
<b>10</b>	<b>benign/negative</b>	<b>1</b>
10	malignant/positive	30

For this reason, the inclusion of outlier values has made the calculation of this ambiguous value range using the minimum-maximum evaluation method over-inclusive, indistinguishable and sometimes impractical for some attributes and some datasets, and some alternatives should be considered.

### 3.5.2 Ambiguous Value Range by Mean Value of Class Labels

There are various methods for dealing with outlier values from a statistical perspective or from a data mining perspective [RL05] [KK17] [WBH02], such as robust regression, Donoho-Stahel estimator, replicator neural networks and the minimum message length (MML) inductive inference process, and they can be implemented in conjunction with the minimum-maximum evaluation process to produce a more reasonable lower and upper boundary for the proposed ambiguous value range.

Instead of implementing one of these elaborate outlier identification and filtering models, this study takes a direct and simplistic approach at this early exploration stage to work out and assign the mean attribute value of the correctly classified negative and positive samples as the lower and upper boundary accordingly, to specify the overlapping area between the center of negative and the center of positive as the proposed ambiguous value range in an arithmetic and simplistic sense.

While calculating this ambiguous value range using a two-class label center approach may sound simplistic, the resulting value range ratio over error ratio in the Pima diabetes dataset seems rather convincing. For example, the ambiguous value range of attribute *plas* is evaluated as only 16% of the overall range value, but the error rate in the value range is 42%; and the ambiguous value range of attribute *mass* is evaluated as only

8% of the overall range value, but the error rate in the value range is 35%. However, other attributes such as *preg* and *age* do not share the same level of success, as the ambiguous value range of *preg* is evaluated as 9% of the overall value range but the error rate in this value range is only 10%, and the ambiguous value range of *age* is evaluated as 10% of the overall value range, but the error rate in this value range is only 16%, as shown in Table 3-18.

Table 3-18 – Pima diabetes dataset’s ambiguous value range by mean value of class label

Second and alternative method of measuring ambiguous value range based on the mean values of TN and TP attribute values						
Attribute	Value range of TN	Value range of TP	Mean value of TN	Mean value of TP	Attribute-error counter (alt)	Value Range ratio : Error ratio
preg	<b>0~13</b>	<b>0~17</b>	3.30	4.87	<b>20</b> (3.30~4.87)	1.57/17 = 9% : 20/200 = 10%
plas	<b>0~175</b>	<b>92~199</b>	109.98	141.26	<b>83</b> (110~141)	<b>31/199 = 16% : 83/200 = 42%</b>
pres	<b>0~122</b>	<b>0~114</b>	68.18	70.82	<b>18</b> (68~71)	<b>3/122 = 2% : 18/200 = 9%</b>
skin	<b>0~54</b>	<b>0~99</b>	19.66	22.16	<b>7</b> (20~22)	2/99 = 2% : 7/200 = 4%
insu	<b>0~485</b>	<b>0~846</b>	68.79	100.34	<b>15</b> (69~100)	31/846 = 4% : 15/200 = 8%
mass	<b>0~57.3</b>	<b>23.3~59.4</b>	30.30	35.14	<b>70</b> (30~35)	<b>5/59.4 = 8% : 70/200 = 35%</b>
pedi	<b>0.08~1.73</b>	<b>0.09~2.42</b>	0.43	0.55	<b>24</b> (0.43~0.55)	0.12/2.34 = 5% : 24/200 = 12%
age	<b>21~81</b>	<b>22~70</b>	31.19	37.07	<b>31</b> (31~37)	6/60 = 10% : 31/200 = 16%

When compared to those supportive findings in the Pima diabetes dataset, the evaluation results from the Wisconsin cancer dataset are less encouraging, and some attributes even show the opposite results. For example, the ambiguous value range of attribute *Clump Thickness* is evaluated as 42% of the overall range value but the error rate in the value range is 67%, and the ambiguous value range of attribute *Single Epithelial Cell Size* is evaluated as 32% of the overall range value, but the error rate in the value range is 49%, so the elevated risk level of error is marginal, as shown in Table 3-19.

Table 3-19 - Wisconsin dataset's ambiguous value range by mean value of class label

Second / alternate method of measuring ambiguous value range based on the mean values of TN and TP attribute						
Attribute	Value range of TN	Value range of TP	Mean value of TN	Mean value of TP	Attribute-error counter (alt)	Value range ratio: Error ratio
Clump Thickness	1~8	1~10	2.96	7.20	<b>30</b> (2.96~7.20)	4.24/10 = 42% : 30/45 = 67%
Uniformity of Cell Size	1~4	1~10	1.33	6.57	<b>35</b> (1.33~6.57)	5.24/10 = 52% : 35/45 = 78%
Uniformity of Cell Shape	1~4	1~10	1.44	6.56	<b>35</b> (1.44~6.56)	5.12/10 = 51% : 35/45 = 78%
Marginal Adhesion	1~6	1~10	1.36	5.58	<b>23</b> (1.36~5.58)	4.22/10 = 42% : 23/45 = 51%
Single Epithelial Cell Size	1~10	2~10	2.12	5.30	<b>22</b> (2.12~5.30)	3.18/10 = 32% : 22/45 = 49%
Bare Nuclei	1~10	1~10	1.35	7.63	<b>23</b> (1.35~7.63)	6.28/10 = <b>63%</b> : 23/45 = <b>51%</b>
Bland Chromatin	1~7	1~10	2.10	5.98	<b>27</b> (2.10~5.98)	3.88/10 = 39% : 27/45 = 60%
Normal Nucleoli	1~8	1~10	1.29	5.86	<b>17</b> (1.29~5.86)	4.57/10 = <b>46%</b> : 17/45 = <b>38%</b>
Mitoses	1~8	1~10	1.06	2.99	<b>3</b> (1.06~2.99)	1.93/10 = <b>19%</b> : 3/45 = <b>7%</b>

However, Table 3-19 also shows that attributes *Bare Nuclei*, *Normal Nucleoli* and *Mitoses* show opposite results. The ambiguous value range of attribute *Bare Nuclei* is evaluated as 63% of the overall range value but the error rate in the value range is 51%; the ambiguous value range of attribute *Normal Nucleoli* is evaluated as 46% of the overall range value but the error rate in the value range is 38%; and the ambiguous value range of attribute *Mitoses* is evaluated as 19% of the overall range value but the error rate in the value range is only 7%. So, at a glance, these three attributes seem to have a lower risk level of error in their ambiguous value range, which is contradictory to what this study has suggested.

When taking a closer look at this result in Table 3-19 and comparing it with the result in Table 3-6, then this contradiction and confusion can be explained by the discrepancy between the two methods used to measure the ambiguous value range of an attribute, one by comparing the maximum and minimum boundary values of class labels and the other by comparing the mean values of those class labels.

### **3.5.3 Discussion on Ambiguous Value Range using Other Methods**

The mixed results from the first two evaluation methods, the overlapping minimum-maximum method and the mean value of class label method, prompted consideration of other evaluation methods, such as the median value of the class label method and the stratified weighted average method, as they all apply one single evaluation to produce a reasonable and reliable ambiguous value range in a systematic way.

A systematic and automatic evaluation of attributes in a dataset can be efficient and effective, but it assumes attributes can be evaluated in the same way within the same algorithm. While it is true that a respectable algorithm can be configured dynamically to suit specific types and volumes of data, and arrays of the parameters and weight can be adjusted accordingly for different attributes during the evaluation process. In essence, the key evaluation logic is still based on the same underlying algorithm even in the case of supervised and combined ensemble evaluation and does not take into account the implicit context between different values within one attribute, and the coherent context between different attributes of every individual sample, or at least not at this stage due to the limitation of technology and artificial intelligence.



It is in this context of individual and contextual consideration, to highlight the importance of understanding the nature and characteristics of individual attributes and their interrelationships in a dataset from a field-specific and domain-driven perspective, that an overwhelming focus and resources on algorithm and automation development may sometimes need to be complemented with specific scenario analysis and case-by-case examination, which is also the case for this ambiguous value range evaluation.

The emphasis of this study is to highlight the possible impact of value range ambiguity and error-sensitive attributes on data classification, and the key objective is to identify and evaluate the potential error-sensitive attributes for stakeholders, and to show that the selection and implementation of a specific evaluation method and algorithm can be important but it is not considered to be the overall priority.

### **3.5.4 Comparing Attribute-Error Counter against Gain Ratio and Information Gain**

Ambiguity can be a source of errors, so ambiguous value ranges of attributes can potentially lead to a higher risk level of classification error. If an attribute has more classification errors in its ambiguous value range, then this attribute may be more prone and more sensitive to error, which is the key reason for this error-sensitive attribute evaluation proposal.

Experiments were conducted to produce the ranking lists of error-sensitive attributes from the UCI datasets, and to compare them against the ranking lists generated by the gain ratio algorithm and information gain algorithm to evaluate their similarity and disparity in order to better understand the value patterns and the dataset as a whole. At the first glance, this three-way comparison appears to be related but inconsistent. For the Pima

diabetes dataset, the three most error-sensitive attributes are also the top three attributes in the gain ratio and information gain ranking lists, as shown in Table 3-20.

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

Rank	Attribute-error counter	Gain ratio	Info gain	Distinct values/Unique samples (WEKA)
1	plas (83)	plas (0.0986)	plas (0.1901)	pedi (517/346 ... 45%)
2	mass (70)	mass (0.0863)	mass (0.0749)	mass (248/76 ... 10%)
3	age (31)	age (0.0726)	age (0.0725)	insu (186/93 ... 12%)
4	pedi (24)	preg (0.0515)	insu (0.0595)	plas (136/19 ... 2%)

For the Wisconsin breast cancer dataset, only *Uniformity of Cell Size* which is the most error-sensitive attribute has been ranked as third in the gain ratio ranking list, meanwhile, *Uniformity of Cell Shape* and *Bland Chromatin*, which are the second and fourth most error-sensitive attributes are not included in the top four positions of the gain ratio ranking list. On the other hand, these same first, second and fourth most error-sensitive attributes have found their matching positions in the information gain ranking list, and only the third ranked attribute is different, as shown in Table 3-21.

Table 3-21 - Top-4 attributes ranked by attribute-error counter, gain ratio and information gain algorithm in the Wisconsin dataset

Rank	Attribute-error counter	Gain ratio	Info gain	Distinct values/Unique samples (WEKA)
1	<i>Uniformity of Cell Size</i> (35)	Normal Nucleoli (0.399)	<i>Uniformity of Cell Size</i> (0.675)	<i>Uniformity of Cell Size</i> (10/0...0%)
2	<i>Uniformity of Cell Shape</i> (35)	<i>Single Epithelial Cell Size</i> (0.395)	<i>Uniformity of Cell Shape</i> (0.66)	<i>Uniformity of Cell Shape</i> (10/0...0%)
3	Clump Thickness (30)	<i>Uniformity of Cell Size</i> (0.386)	Bare Nuclei (0.564)	Bare Nuclei (10/0...0%)
4	<i>Bland Chromatin</i> (27)	Bare Nuclei (0.374)	<i>Bland Chromatin</i> (0.543)	<i>Bland Chromatin</i> (10/0...0%)

These similar ranking patterns of highly error-sensitive attributes between the Pima Indian diabetes dataset and the Wisconsin dataset can possibly be explained by the distinct values and unique samples statistical function in WEKA, as shown in the column “Distinct values/Unique samples (WEKA)” in Table 3-20 and Table 3-21, as well as Table 3-3 and Table 3-7. According to WEKA’s user manual, the distinct function returns “the number of different values that the data contains for this attribute”, and the unique function returns “the number (and percentage) of instances in the data having a value for this attribute that no other instances have” (WEKA manual for version 3-7-4, 2011, p.46), which means, the number of samples that have only one such attribute value once and uniquely, no other sample shares the same attribute value in the dataset.

To be specific, the top three attributes in the Pima Indian diabetes dataset - *plas*, *mass* and *age*, have their distinct value and unique sample counts also in high ranking positions by “Distinct values/Unique samples (WEKA)” – fourth for *plas* (136/19 ... 2%), second for *mass* (248/76 ... 10%) and fifth for *age* (52/5 ... 1%), as shown in Table 3-3 and Table 3-20, which means they have a relatively high number of distinct attribute values and unique records. If these higher differential distinct values and unique ratios can partially explain the higher info gain and gain ratio rankings according to their theoretical definitions [Sha48] [Qui93], then the replication of the distinct values into attribute-error counters and unique ratios into the ratio of error samples over the total samples within the specified ambiguous value range, as shown on the left-hand side of Table 3-22 with higher attribute-error counters and a higher ratio of error samples over the total samples within the specified ambiguous value range, may also help to explain and support this proposal

of error-sensitive attribute evaluation and the supportive results in the Pima diabetes dataset.

However, the results for the distinct value and unique sample counts for the Wisconsin cancer dataset appear to be less meaningful than the gain ratio and information gain ranking functions, as shown in Table 3-7 and Table 3-21. One logical way to explain this confusion is the fact that all the involved attribute values have been normalized to the same scale between one and ten which limits them to ten integer values by the original data provider, and this might have subsequently equalized those distinct and unique values in between each of the ten integer values, leading to less meaningful distinct and unique results in WEKA.

Table 3-22 - Comparison of attribute-error count ratio between the Pima and Wisconsin datasets

Pima Diabetes Dataset			Wisconsin Cancer Dataset		
Rank	Attribute-error counts	Ratio to all samples in ambiguous value range	Rank	Attribute-error counts	Ratio to all samples in ambiguous value range
1	plas (83)	31% (83/ <b>267</b> ... 768)	1	Uniformity of Cell Size (35)	18% (35/ <b>194</b> ... 699)
2	mass (70)	33% (70/ <b>214</b> ... 768)	2	Uniformity of Cell Shape (35)	16% (35/ <b>223</b> ... 699)
3	age (31)	34% (31/ <b>92</b> ... 768)	3	Clump Thickness (30)	8% (30/ <b>375</b> ... 699)
4	pedi (24)	27% (24/ <b>90</b> ... 768)	4	Bland Chromatin (27)	11% (27/ <b>239</b> ... 699)
5	preg (20)	29% (20/ <b>68</b> ... 768)	5	Marginal Adhesion (23)	13% (23/ <b>172</b> ... 699)
6	pres (18)	32% (18/ <b>57</b> ... 768)	6	Bare Nuclei (23)	19% (23/ <b>119</b> ... 699)
7	insu (15)	21% (15/ <b>72</b> ... 768)	7	Single Epithelial Cell Size (22)	14% (22/ <b>159</b> ... 699)
8	skin (7)	18% (7/ <b>39</b> ... 768)	8	Normal Nucleoli (17)	15% (17/ <b>117</b> ... 699)
			9	Mitoses (3)	9% (3/ <b>35</b> ... 699)

Instead of using the distinct value and the unique sample counts function in WEKA, the ratio comparison of the attribute-error counter value over the count of all instances within the ambiguous value range is used. The comparison results for the Wisconsin cancer

dataset are reported in Table 3-22, confirming that the two most error-sensitive attributes, *Uniformity of Cell Size* and *Uniformity of Cell Shape*, also have rather high counts of instances inside their respective ambiguous value ranges: *Uniformity of Cell Size* has 194 from 699 samples and *Uniformity of Cell Shape* has 223 from 699 samples. Also, they both have a higher ratio of attribute-error counter value over the count of all instances within the ambiguous value range: *Uniformity of Cell Size* is 18% (35/194) and *Uniformity of Cell Shape* is 16% (35/223) when compared to the other seven attributes.

The higher attribute-error counter values can be simply a reflection of the fact that more samples have their attribute values inside their ambiguous value ranges, so statistically, there are more instances within the range and more chance for errors. This reasoning is a good explanation for the Pima dataset, but it cannot explain why the two attributes with the second and fourth most error-sensitive values in the Wisconsin dataset, *Clump Thickness* with 375 and *Bland Chromatin* with 239, are ranked lower than *Uniformity of Cell Size* with 194 and *Uniformity of Cell Shape* with 223. So, this situation indicates that having more samples in the ambiguous value range does not necessarily lead to a higher attribute-error count automatically in a proportional way, and there can be other factors involved.

While this discussion on the ratio of error samples over total samples within the ambiguous value range has prompted the consideration of using the error density function as another alternate evaluation method for error-sensitive attributes, this alternative is discussed in Section 3.6 Related Work, but is outside the current development scope.

### 3.5.5 Mixed Results from Potential Error-Reduction Measure

Each attribute is supposed to have its own risk level of classification error in a dataset. If an attribute has a higher risk level of error, it may subsequently have a more significant impact on the classification process and result. In the experiments with the UCI datasets, the resulting error-sensitive attribute ranking order shows a certain level of conformity with the gain ratio and info gain ranking order, as discussed in the previous sessions. The classification processes are re-run to test the validity of such a claim of error-sensitivity level by filtering out the most error-sensitive attributes as a potential error-reduction measure and some mixed results have been produced.

In the Pima dataset with the filtering out of the top two most error-sensitive attributes, the accuracy rate in its re-run of the classification process has decreased from 73.96% to 67.06%, as shown in Table 3-4. This result may look discouraging at first; however, a closer look at the earlier three-way comparison analysis shows that the two most error-sensitive attributes, *plas* and *mass* are also the two highest ranked gain ratio attributes. Because the gain ratio method is the underlying algorithm implemented as the WEKA's J48/C4.5 decision tree classification model, when these two most significant attributes are filtered out and removed from the classification process, this likely has a negative impact on the model's performance and accuracy rate, therefore resulting in the decrease from 73.96% to 67.06%.

On the other hand, after filtering out the two most error-sensitive attributes, *Uniformity of Cell Size* and *Uniformity of Cell Shape*, the accuracy rate in its re-run of the classification process for the Wisconsin breast cancer dataset improves from 94.13% to

95.71%, as shown in Table 3-8. This improvement could be because the two most error-sensitive attributes, *Uniformity of Cell Size* and *Uniformity of Cell Shape*, are not the two highest ranked attributes by gain ratio, subsequently, their removal has no major impact on the J48/C4.5 decision tree classification model which is based on the gain ratio algorithm. Because these two most error-sensitive attributes are filtered out, this likely leads to the reduction of misclassification risk. In the meantime, the significant gain ratio associated attributes still remain for the J48/C4.5 decision tree's processing logic, therefore the classification accuracy increases from 94.13% to 95.71%. This improvement may not look significant, but it is a supportive sign for the proposed idea of error-sensitive attribute evaluation, an encouragement to explore further the topic of ambiguous and error-sensitive attributes and value patterns at a later stage.

One hypothesis on error-sensitive attributes which arises from the experiment analysis using the Pima diabetes dataset and the Wisconsin dataset is, if the most error-sensitive attributes are not the most significant attributes that are used by the underlying classification algorithm, then the removal or weighting down of these specific error-sensitive attributes can help enhance the classification performance, and the Wisconsin dataset experiments appear to support this first point of the hypothesis.

If the most error-sensitive attributes are also the most significant attributes for the underlying classification algorithm, then the removal or weighting down of these error-sensitive attributes can actually hinder rather than improve the classification process, and the Pima dataset experiments appear to support this part of the hypothesis. On the surface, this hypothesis appears to explain the mixed results when experimenting with the Pima

diabetes dataset and the Wisconsin dataset, but in reality, finding sufficient evidence to prove a hypothesis is not always so simple.

### 3.5.6 Follow-up Tests on the Error-Sensitive Attribute Hypothesis

Three other UCI datasets are used in the follow-up experiments to test the above hypothesis. In the Ecoli dataset experiment, the most error-sensitive attribute *chg* identified by the evaluation process is ranked as the least significant attribute by both the gain ratio and information gain method, and the removal of this attribute does not have any effect on the re-classification process or the result. According to the attribute information provided by the dataset owner, *chg* is a binary attribute to indicate “the presence of charge on N-terminus of predicted lipoproteins” and uses the value of 1 and 0.50 to indicate whether it is present or not.

A closer look at the data reveals that only one sample has a value of 1, all other 335 samples have a value of 0.50, which means the attribute itself is not an effective measure in the classification process at all, explaining why both the gain ratio algorithm and information gain algorithm rank *chg* as the least significant attribute because of its non-differentiating nature. This also explains why the proposed evaluation process incorrectly ranks it as the most error-sensitive attribute because of the severely distorted value distribution in attribute *chg*, where 335 samples have a value of 0.50 and 17 of the 18 classification errors, and only one sample has a value of 1 and is also one of the 18 classification errors in this test.

This test case of attribute *chg* highlights one major issue with the proposed error-sensitive attribute evaluation method in its current form. When values of an attribute are



heavily skewed or in an extremely polarized state, this attribute may not be a suitable candidate for the proposed evaluation process.

On the other hand, *alm1* and *lip* are evaluated as the second and third most error-sensitive attributes, and they are also ranked as the first and third by the gain ratio algorithm underpinning the J48/C4.5 classifier, so the filtering out of these two attributes may help reduce the risk level of classification errors, however, this may also decrease the classification accuracy rate on a bigger scale. Indeed, the re-run classification accuracy rate decreases from 94.64% to 92.56% after filtering out *alm1* and *lip*, as shown in Table 3-10, while the accuracy rate remains unchanged after filtering out *chg*. From a speculative perspective, the experiment results on the Ecoli dataset appear to support the latter part of the error-sensitive attribute hypothesis.

In the Liver Disorders dataset experiment, the most error-sensitive attribute *sgot* identified by the proposed evaluation is ranked the least significant attribute by both the gain ratio and information gain algorithms, which is similar to the situation of attribute *chg* in the Eocli dataset as just discussed. However, the filtering out of attribute *sgot* improves the accuracy rate from 67.54% to 68.70% in the re-run of the classification process, which is different from the Eocli dataset experiment result. When the second most error-sensitive attribute *mcv* which is ranked the third most significant by both the gain ratio and information gain algorithms, is filtered out and the classification process is re-run, the accuracy improves further to 70.43%, which is an almost 3% improvement on the original classification result. These results appear to support the first part of the hypothesis.

In the Page Blocks dataset experiment, *mean\_tr* is the most error-sensitive attribute, and it is also ranked the third most significant attribute by the gain ratio algorithm and the second most significant attribute by the information gain algorithm. The filtering out of this single attribute leads to a slight decrease in the classification accuracy rate from 97.22% to 97.20%.

This marginal decrease in the accuracy rate can be related to the latter part of the hypothesis, similar to the test cases with the filtering out of attributes *alm1* and *lip* in the Ecoli dataset, as well as the filtering out of attribute *plas* and *mass* in the Pima diabetes dataset. On the other hand, the filtering out of the second and third most error-sensitive attribute, *p\_black* and *eccen* improved the accuracy rate by a miniscule margin from 97.22% to 97.26%, and this improvement can be explained by referring to the first part of the hypothesis. Because of these mixed results, it is fair to say that these test cases on the page block dataset experiment demonstrate a certain level of support for both the first and the latter part of the error-sensitive attribute hypothesis. While the level of certainty is variable and arguable, the collective and supportive evidence from these UCI datasets is explainable and undeniable.

### 3.5.7 Comparison with Experiments by Other Researchers

As outlined in the beginning of this chapter, the key objective of this study is not about algorithm development for error reduction or classification performance enhancement, it is to highlight the issue of the ambiguous value ranges of the attributes and their associated risk level in relation to classification errors, to explore and evaluate error-sensitive attributes and their potential impact on the classification process and results for the purpose of

gaining a better understanding of error-sensitive value patterns and the dataset in a data-centric way.

Therefore, another way to analyze the proposed evaluation and its experiments is to compare studies by other researchers on the same datasets with a focus on data-centric analysis rather than algorithm development, to check for any special mention of the identified error-sensitive attributes, to seek supporting evidence from direct or implicit references recognized by other researchers that may shed light on similar attribute evaluation ideas.

Finding such data-centric references for these specific datasets is easy, but understanding domain-specific terms and technical jargon is not so easy. The first associated research paper [SED88] mentioned within the Pima diabetes data file focuses on the ADAP neural network model and adaptive learning algorithm, together with the background information about the Pima diabetes sample selection and the introduction of the eight attributes, especially the Diabetes Pedigree Function which is based on a genetic relationship in terms of family members and their diabetes mellitus history. Despite the inclusion of some detailed data-centric information, the paper gives no specific analysis of classification errors and their possible relationship with specific high-risk attributes and potential ambiguous value ranges.

A good number of other data classification models have been developed and used in experiments based on a detailed analysis of this Pima diabetes dataset. For example, the ARTMAP-Instance Counting algorithm [CM98] was applied in a medical diagnostic study on four medical datasets, the Pima diabetes dataset; the hybrid Bayes classifier and

evolutionary algorithm [RDK01], the automated comparative disease profile generation and diagnosis system [WA04], and the general regression neural networks model [KY03]. Although they introduced new developments to classification algorithms and made enhancements at various levels using the Pima diabetes dataset for testing and in-depth result analysis, there is no worthwhile discussion or analysis on possible causes of classification errors and their association with specific high-risk data attributes and potential ambiguous value ranges.

Similarly, in relation to the Wisconsin cancer dataset, of the data classification models that have been studied, such as the multi-surface pattern separation model [WM90], the two parametric bilinear programming methods for feature minimization models, one with bounded accuracy and one with limited feature minimization based on best linear discriminant [BB98], and the cancer image analysis application with a diagnosis and prognosis capability based on two other linear programming methods [MSW94], the focus of these experiments with the Wisconsin cancer dataset is primarily on classification performance and a comparison between their models and other benchmark models, such as CART [BFO84] and C4.5, and there is no specific discussion on the possible role of and specific impact of certain value ranges within certain attributes with a higher level of risk in relation to classification errors.

It is noted that one prognosis classification method [MSW94] [SMW95] indicates that “the best predictive models are found using just the nuclear features, and are not improved by the traditional medical prognostic factors of tumor size and lymph node status” (Street, Mangasarian & Wolberg, 1995, p.7), however, there is no further discussion on

what specific data elements and attributes would be more likely associated with the reported errors, such as which attribute and what value range may have a potential association with “the prognostic system [which] has an expected error of 13.9 to 18.3 months” (Mangasarian, Street & Wolberg, 1994, p.1).

### **3.6 Related Work**

In addition to the individual development and analytic work on this proposal of an error-sensitive attribute evaluation model based on the suggested concept of ambiguous value range and attribute-error counter, and the many experiments that were conducted to test it against other attribute ranking lists generated by some reputable attribute selection algorithms, such as the gain ratio algorithm and the info gain algorithm, a significant amount of time and effort has also been spent on reading and reviewing the literature on related topics.

#### **3.6.1 Study of Ambiguous Data**

Ambiguity is usually studied and associated with philosophy and politics, linguistics and graphics, and is a less popular topic for study in terms of data mining and data classification. When ambiguity is studied in association with data mining and data classification, the focus of study tends to aim at value patterns at a sample level or towards outlier values and is primarily from a data preprocessing and data cleaning perspective.

A classification scheme has been proposed by Trappenberg and Back [TB00] to detect and re-classify ambiguous data based on the presumption that ambiguous data are closely spaced and may not be separable during classification, and with some attribute

value patterns belonging to more than one class label, therefore leading to overlapped value patterns, no clear class label decision boundary and a higher level risk of false classification. The detection part of the scheme is based on a k-NN algorithm to check if a data point has sufficient neighboring points out of k selection to be of one of the class labels. If the majority of the neighboring points belong to one particular class label, then this same class label will be assigned to the targeted data point; if no majority can be decided, then this targeted data point will be assigned to the ambiguous data group. After all data points are tested, the established ambiguous data group will be re-classified and used as a part of the training samples to re-classify the test data, and as no specific classifier is preferred under this scheme, a feedforward neural network classifier is used for testing.

The classification scheme with ambiguous data developed by Trappenberg and Back has by far the closest similarity with the proposed error-sensitive attribute evaluation model in terms of its approach. The researchers state in their introduction, “the basic idea is quite straightforward: rather than seek a complex classifier, we aim to first examine the data with the aim of removing any ambiguities” (Trappenberg & Back, 2000, p.1), and include in its cautionary conclusion, “hence one should consider avoiding predictions of some data in areas which are highly ambiguous ... when predictions have to be made with particular caution.” (Trappenberg & Back, 2000, p.5)

However, this classification scheme is designed as a pre-processing routine with a focus on individual data points to identify and isolate ambiguous data as a new input for its own round of classification, which is significantly different to the proposed evaluation

model as a post-classification routine and with a focus on attributes to identify and isolate attributes with a higher level of errors from ambiguous value ranges.

A follow-up study on ambiguous data by Trappenberg and Hashemi [HT02] adopted the Coverage versus Performance (CP curve) function as a way to identify and isolate ambiguous data from a dataset and adopted the support vector machine (SVM) as its specific classifier by utilizing and varying the regularization parameter to generate the CP curve and ambiguous data group. This study found that in some experiments, even the removal of a small portion of ambiguous data had a relatively significant accuracy improvement, an indication that ambiguous data can sometimes have a rather negative impact on the classification process.

However, in some other experiments, there is not much difference between the classification accuracy on the original dataset and the cleaned data after the removal of ambiguous values, and the authors of the study explain such “counter-intuitive” results by highlighting one advantage of SVM’s support vector processing logic which is based on data points with dominant contributions by maximized margins, therefore ambiguous data may not be included in key label determining steps.

Motivated by the k-NN and SVM approaches on ambiguous data identification, and by associating ambiguous data with uncertain and overlapping feature values, another SVM variation in the “shape of sphere” has been developed [LWN06]. This sphere classification model incorporates the sequential minimal optimization algorithm using SVM with Gaussian kernels to construct classification rules in the feature space of a dataset to cover data points under the shape of two spheres, one sphere for the positive samples and

one sphere for the negative samples. Samples located in the overlapping area of the two shapes are subsequently determined to be ambiguous data and samples located outside the two spheres are determined to be the outliers.

Instead of using two spheres to identify ambiguous values in the overlapped area, a two-neural-network model has recently been developed to identify ambiguous data [LM18]. In this new model, the first neural network computes the average predicted values from certain input of the dataset, and the second neural network evaluates the output from the first one and generates the expected error (or variance) based on that output; subsequently, the specific input that is associated with such errors can be classified as ambiguous values.

The primary focus of ambiguous data analysis in terms of data mining and classification has been on algorithm and application development, like the various enhancements to the f-NN algorithm, the SVM algorithm and neural networks that have been outlined above. Their related theories are mostly based on statistics, such as entropy and information theory [Sha48] [Qui93], prior and posterior probability [TBA99] [PF01], and variations of uncertainty theories [LM18] [Kni21] [KMM05], and they have been widely studied by the research community and by various industries.

### **3.6.2 Comparing Attribute Selection**

This study suggests that each attribute may have its own level of risk leading to classification errors and evaluates the effectiveness of such error-sensitive attribution identification by comparing top error-sensitive attributes ranked by attribute-error counters against



the most significant attributes ranked and selected by other reputable algorithms, such as the gain ratio algorithm and the information gain algorithm.

The information gain algorithm is based on the concepts of entropy and information gain described in Shannon's information theory [Sha48]. While the information gain algorithm has been an effective attribute selection method, it also has a strong bias for selecting attributes with many different outcomes, for example, the trivial attribute of Record ID or Patient ID, which is not supposed to be related with the classification process itself.

To address this bias deficiency, information gain ratio was developed as the attribute selection criterion in C4.5 to apply "a kind of normalization in which the apparent gain attributable to tests with many outcomes is adjusted" (Quinlan, 1993, p.23), which means the information gain is to be divided into as many possible partitions or outcomes for an attribute, so it "represents the proportion of information generated by the split that is useful, i.e., that appears helpful for classification" (Castillo, 2012, p.109), whereas the gain ratio method selects an attribute which can maximize its gain ratio value [Qui93].

There are a variety of other attribute ranking and selection methodologies that have also been developed over the years, such as the Gini index method that was built into the CART decision tree model, the RELIEF algorithm, the sequential forward selection (SFS) method and the sequential backward elimination (SBE) method [BFO84] [Alp09] [HK06] [SIL07] [WF05], as well as some newly established techniques such as SAGA [GS10] and UFSACO [TMA14] which were developed as a pre-process procedure or as an integrated part of the classification process. One key logic shared by these algorithms is to select and prioritize the most informative and differentiating data attributes before or during

classification [LMS10], as shown in Figure 3.4, which is different to the approach adopted in this proposed study which is to rank and evaluate attributes as a part of the post-classification analytic process.

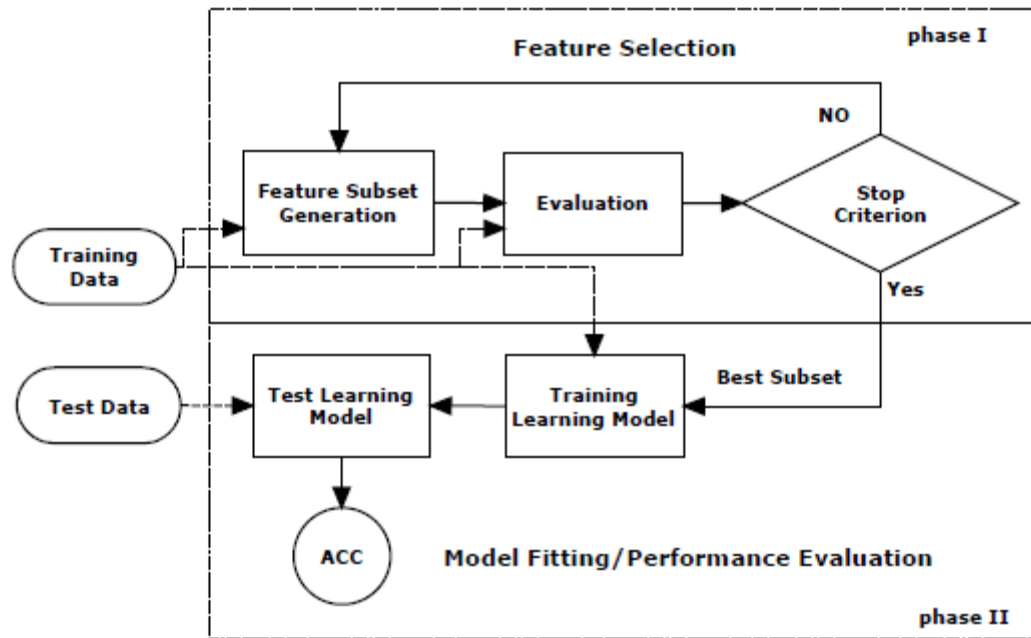


Figure 3.4 - A typical feature ranking and selection model for data classification

Because the focus of this chapter is on ambiguous value ranges and error-sensitive attribute evaluation, further details on the related work on attribute selection are discussed in the next chapter in relation to decision tree branch evaluation.

### 3.7 Summary

In a schema of pursuing the further understanding of data rather than following the common track of algorithm development, to address the question raised in the start of this chapter - “Are there specific attributes in a dataset which may be more vulnerable to cause ambiguity and errors, and if yes, can they be identified in a systematic way?”, an error-

sensitive attribute evaluation model has been proposed as part of the answers by defining and adapting three data-centric terms, ambiguous value range, attribute error counter and error-sensitive attribute, and the initial experiments demonstrated a number of supportive test cases. While the first stage of this study into ambiguous value range and error-sensitive attribute evaluation can now be concluded, the limitation on number of evaluation methods used for comparison and the number of datasets used in testing and result validation should also be acknowledged.

Nevertheless, such limited yet encouraging test results led to the development of an error-sensitive branch identification process for the decision tree classification model to further examine the potential correlation between the error-sensitive value patterns and classification results in a systematic way, to expand the analytic scope from an attribute level to a more specified value pattern level which may reveal more relevant attributes and with specific value pattern association, which is the focus of study in Chapter 4.

## CHAPTER 4

### **Exploration of the Weakest and Error-Sensitive Decision Branches**

To demonstrate and report one's understanding about data in an effective way, systematic exploration and classification methodologies have been developed. Decision trees have a reputation of being efficient and illustrative in classification learning and the majority of the research effort has been focused on making classification improvement in a head-on style with a wide-range of research topics, such as tree algorithm development and refinement, attribute selection and prioritization, sampling technique improvement, and the addition of a cost matrix and other performance-enhancing factors.

One less commonly studied topic concerns the characteristics of classification errors and how they may be associated with specific attributes due to correlation or causation, and what particular value patterns are more likely linked to such specific associations of classification errors. A study into the potential correlation between error-sensitive attributes and classification error has been outlined in Chapter 3, a further study on how specific error-sensitive value patterns involving multiple attributes and specific value ranges may play a role in classification errors is discussed in this chapter, and this discussion is based on the exploration of the most error-sensitive value patterns in the form of the weakest and error-prone decision paths and branches by adopting a decision tree classification model as its processing platform.

#### **4.1 Overview of Decision Trees and Error-Sensitive Pattern Identification**

The formation of each decision tree branch is different due to the variation of attributes and their split-point values at various node levels, subsequently, the contents and conditions of

individual decision rules are different in the form of different tree branches, therefore, the classification result from each decision rule and tree branch may be different. Some branches may have a wider inclusion of sample records than others, and some branches may have more effective and significant decision paths which may result in higher accuracy rates than others. This consideration has led to some practical questions about a decision tree classifier:

- How to create a new tree model or modify the current one for better performance?
- What are the most influential attributes used in a decision tree classification task?
- What other factors, such as feature-selection routine and cost-sensitive matrix, can be added to enhance the decision tree classification performance?

These are some of the forward-thinking questions commonly asked during post-classification analysis. Some less forward-thinking questions are:

- Which tree branch is the weakest and associated with more misclassification errors in a decision tree?
- What attributes and their values are more likely to cause or be associated with classification errors?

The study of the weakest branches of a decision tree can relate to the tree pruning topic. One typical way to handle weak branches is pruning and a good number of decision tree pruning methodologies have been developed over the years, such as the error reduction pruning method [Qui93], the cost-complexity pruning method [BFO84], and the minimum error pruning method [CB91]. This study is not about creating a new pruning technique, but to explore an alternative way to identify and evaluate the weakest branches in a decision

tree, to highlight the specific error-sensitive patterns to stakeholders as a way of gaining further understanding about the data, and the potential correlation between certain value patterns and classification errors, within the context of adjacent and associate decision rules in the form of neighboring branches and nodes for a better understanding of the data and specific patterns.

The starting point of this exploration process is not before, not during, but after the fact, that is, after the pre-determined feature selection and sub-tree pruning routine have already been completed as part of the classification process, so the exploration can examine the overall classification result in a retrospective way, as well as having a particular focus on the most error-prone decision tree branches and possible relationship patterns between these branches and the most error-sensitive attributes.

This proposed exploration is carried out in connection with the evaluation process of error-sensitive attributes, a proposal which has been discussed in Chapter 3. This evaluation proposal suggests three specific terms, ambiguous value range, attribute-error counter and error-sensitive attribute, to describe how the most error-sensitive attributes can be identified by ordering the attributes' risk level which is based on each attribute's error count within its ambiguous value range. In this proposed exploration, two new routines have been added onto the mentioned evaluation process. The first routine is to rank the resulting decision tree branches according to their associated predicted error counts, and the second routine is to examine the existence of the most error-sensitive attributes and their value ranges on the identified risky branches. While an initial report has summarized

parts of this exploration proposal [Wu13], more analysis on the experiments and results are now included in the next few sections.

Here is a brief scenario to outline a practical issue. In a binary data dataset with 10,000 sample instances and 50 data attributes, a classification task is performed by a decision tree model enabled with its own feature selection and pruning routine. The result is a pruned tree with 60 nodes, 30 leaves and 100 misclassification errors. Of the 30 root-to-leaf branches, some have higher error counts than others. Our research question based on this scenario is, what are the potential correlation patterns between the most error-associated branches and those error-sensitive attributes and value ranges?

Figure 4.1 is an example of a decision tree with its size as seven, which means, three nodes and four root-to-leaf branches. In this study, the terms, *tree branch* and *branch* refer to a full branch from the tree root to an end leaf. A portion of such a branch or a sub-tree structure is outside the scope of this study at this stage.

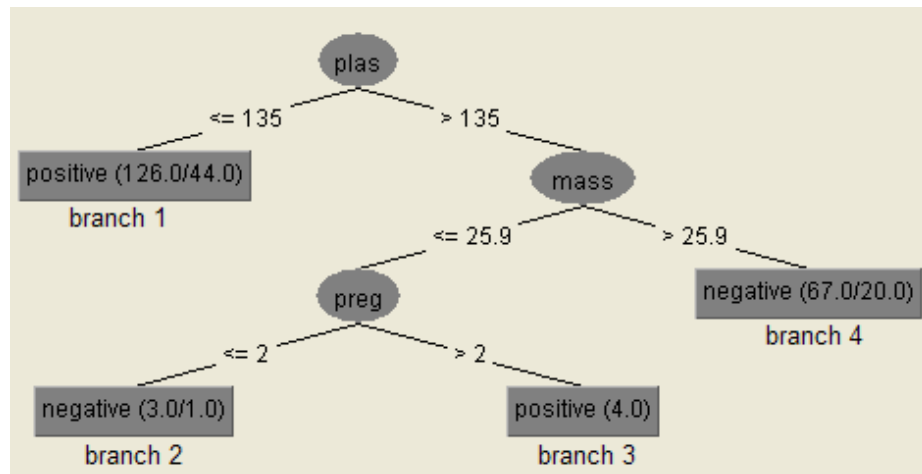


Figure 4.1 - A sample tree with each branch showing its predicted error-rate.

During the course of this exploration and study, it is acknowledged that even when potential correlation between a tree branch and classification errors can be identified, it is not confirmation that the associated attributes and value ranges on that branch are the sole reason for the cause of the errors; but rather, may highlight and imply a possible connection between the identified value patterns with a higher risk level of classification errors, to help raise the necessary attention about such patterns and connections to the stake-holders and to serve as motivation to understand the data further.

The rest of this chapter is organized as follows. Section 4.2 describes the process details of this weakest tree branch exploration. Section 4.3 summarizes the experiments on five datasets. Section 4.4 compares and analyzes the experiment results, Section 4.5 reviews some early influential works that have inspired and guided this exploration and study task, and Section 4.6 concludes this stage of the exploration progress and outlines a possible development plan for future study.

## **4.2 Exploring Decision Trees for Error-Sensitive Branch Evaluation**

This error-sensitive tree branch exploration process can be considered as a new extension to the error-sensitive attributes evaluation process discussed in Chapter 3. These two processes can work together as a part of the post-classification analysis to identify and evaluate error-sensitive attributes and tree branches in an attempt to examine the potential correlation between the identified value patterns and classification errors, as shown in Figure 4.2.

Figure 4.2 illustrates that the new exploration process highlighted in blue on the left-hand side path, can run in parallel with the proposed error-sensitive attribute evaluation process on the right-hand side, which is based on Figure 3.2, to examine the classification



results and with a special focus on errors and their potential correlation with error-sensitive attributes and the weakest tree branches, to help develop a further understanding about the data and a more effective error-reduction measure.

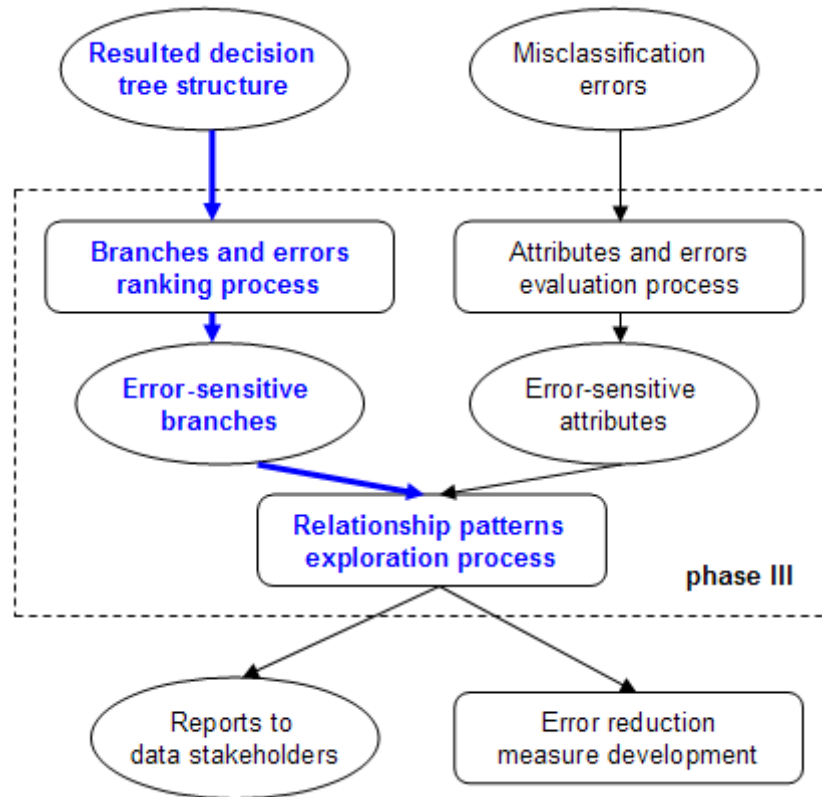


Figure 4.2 – Tree Branch Exploration is added to phase-III of the classification process

On completion of a typical classification task, this post-classification exploration process begins. The initial ranking process for the resulting decision tree branches starts after the evaluation process for error-sensitive attributes to utilize the identified error-sensitive attributes in a later step of the weakest branch exploration.

The key task of the branch ranking process is to convert all root-to-leaf branches from the left-most branch to the right-most branch into a set of decision rules with their exact attributes and split-point value details, as well as the associated error count of each

branch. These error counts are then sorted in order, and the branches with the highest error counts can subsequently be identified. The highly error-sensitive branches can then be compared with the most error-sensitive attributes identified by the evaluation process proposed in Chapter 3. An overview of this exploration process is outlined in the following step-by-step summary.

<p><b>Input:</b> (1) Initial classification result in the form of a decision tree  (2) Original dataset with their newly classified result labels</p> <p><b>Output:</b> (1) A list of branches in the form of decision rules ranked by their error rates  (2) A list of attributes ranked by their attribute-error counter values  (3) Identification of possible weakest tree branches and associated error-sensitive attributes and value ranges on these branches</p> <p><b>Exploration process in six steps:</b></p> <p>Step-1: Serialize and convert the resulted tree structure into its corresponding set of classification rules, and each rule is attached with its error rate from initial classification run</p> <p>Step-2: Sort tree branches by their error rates in order to identify the potentially weakest tree branches with higher error rates</p> <p>Step-3: Compute ambiguous value range and attribute-error counter for each attribute</p> <p>Step-4: Sort attributes by their attribute-error counter values and select the most error-sensitive attributes</p> <p>Step-5: Explore and highlight error-sensitive attributes and value ranges on potentially weakest tree branches ... by going through:</p> <ul style="list-style-type: none"> <li>for each identified weakest tree branch <ul style="list-style-type: none"> <li>for each identified highly error-sensitive attribute <ul style="list-style-type: none"> <li>search current error-sensitive attribute from the current branch/decision rule</li> <li>if current attribute is found in the current rule <ul style="list-style-type: none"> <li>compare its current split-point value with its ambiguous value range</li> <li>if its current split-point value is within its ambiguous value range <ul style="list-style-type: none"> <li>then highlight this attribute and split-point value on the branch/decision rule</li> </ul> </li> </ul> </li> </ul> </li> </ul> </li> </ul> <p>Step-6: Examine the identified error-sensitive branches and associated error-sensitive attribute and value patterns to help develop further understanding about the data and error-reduction measure</p>
---

To verify whether this exploration process can actually identify any highly error-sensitive tree branches and specific relationship patterns between classification errors,

error-sensitive attributes and tree branches, experiments have been performed on a number of real-world datasets, and their results are analyzed and discussed in the next two sections.

### 4.3 EXPERIMENTS

Another five datasets from the UCI Machine Learning Repository have been used to test this exploration idea, and the decision tree classifier for testing is the J48 tree model in WEKA, which is based on the reputable C4.5 algorithm, with all the standard system configurations, including the stratified 10-fold cross-validation as the training and testing option.

On completion of the initial classification process and as part of the post-classification analytic tasks, the resulting decision tree structure is serialized and converted into a set of classification rules and each rule is attached with its own error rate resulting from this particular decision rule. These tree branches in the form of decision rules are subsequently sorted by their classification error rates and the highly ranked tree branches are then checked for any error-sensitive attributes and associated split-point values.

The five datasets used in these experiments are different to the five datasets reported in Chapter 3 and they show supportive results in a variety but limited number of test scenarios, and various error-sensitive attributes can be identified in all the weakest tree branches. When testing the error-reduction measure by filtering out the potentially error-sensitive attributes, their re-classification results present a more consistent improvement pattern when compared to the mixed results produced by the error-sensitive attribute experiments reported in Chapter 3 using the five other datasets; and in one test case of the

Glass ID dataset, its accuracy rate increased by 2.8%. Some of the experiment results are summarized as follows.

#### 4.3.1 The Connectionist Bench (Sonar, Mines vs. Rocks) Sonar Dataset

This binary dataset has two class labels, M for Metal cylinder and R for Rock object, and it contains 208 samples and 60 attributes, each attribute representing a particular sonar frequency band, and values within an attribute represent the bouncing energy strength value of that particular sonar frequency. All have been normalized to the range between 0.0 and 1.0. In the dataset, 111 samples are determined as M and 97 samples are determined as R [GS88].

The initial classification process correctly classifies 148 samples with an accuracy rate of 71.15%, and 60 samples are misclassified. The resulting decision tree has 35 nodes and 18 branches, as shown in Figure 4.3.

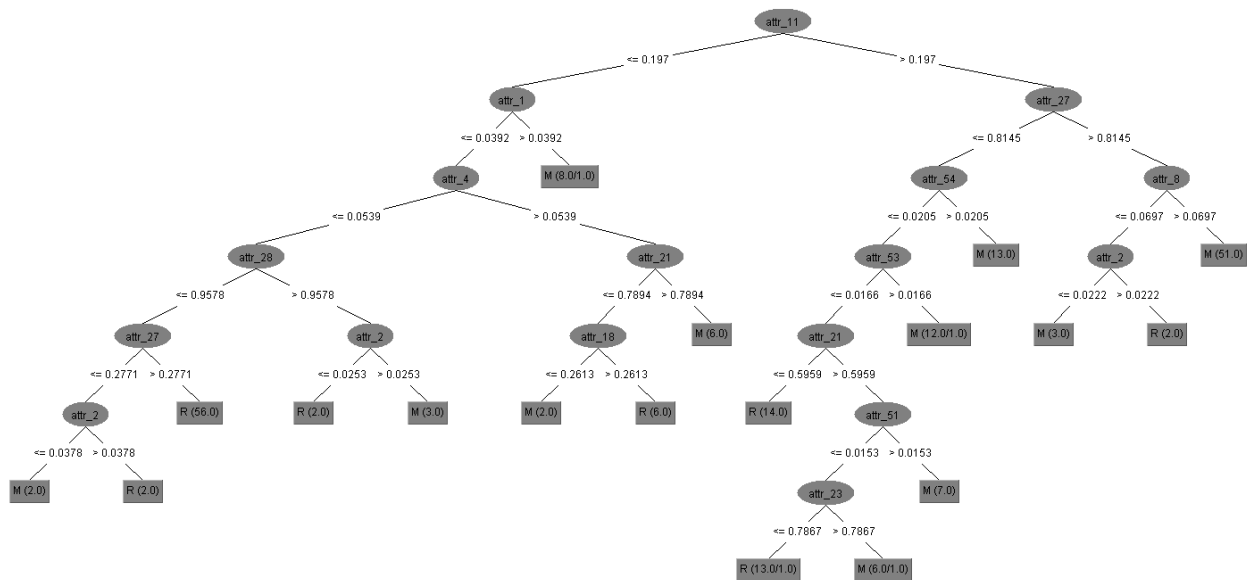


Figure 4.3 - Decision tree from the Connectionist Bench sonar dataset

On completion of the error-sensitive attribution evaluation process, attributes *attr\_11*, *attr\_48*, *attr\_45* and *attr\_9* are ranked as potentially the four most error-sensitive sonar frequency bands. When comparing the ranking list generated by the gain ratio algorithm and the information gain algorithm, the top ranked attribute *attr\_11* is also ranked as the most significant frequency band by both the gain ratio and the information gain, and the fourth ranked *attr\_9* is ranked third by both the gain ratio and the information gain, as shown in the upper sub-table in Table 4-1. It is interesting to note that both the gain ratio and the information gain rank the same top three significant attributes but differed in the fourth ranking position.

Table 4-1 – Exploration of error-sensitive branches in the Sonar dataset

Rank	Attribute-error counter	Gain ratio	Info gain		
1	<i>attr_11</i> (26): 0.17~0.29	<i>attr_11</i> (0.2053)	<i>attr_11</i> (0.2014)		
2	<i>attr_48</i> (24): 0.07~0.11	<i>attr_12</i> (0.1803)	<i>attr_12</i> (0.1779)		
3	<i>attr_45</i> (23): 0.14~0.25	<i>attr_9</i> (0.1634)	<i>attr_9</i> (0.1498)		
4	<i>attr_9</i> (21): 0.14~0.21	<i>attr_44</i> (0.1589)	<i>attr_10</i> (0.143)		

Rank	Decision tree branch (classification rule)	Error count	Sample count	Error rate
1	<i>branch 12</i> : <i>attr_11</i> > 0.197 and <i>attr_27</i> <= 0.8145 and <i>attr_54</i> <= 0.0205 and <i>attr_53</i> <= 0.0166 and <i>attr_21</i> > 0.5959 and <i>attr_51</i> <= 0.0153 and <i>attr_23</i> > 0.7867: M	1	6	16.67%
2	<i>branch 9</i> : <i>attr_11</i> <= 0.197 and <i>attr_1</i> > 0.0392: M	1	8	12.50%
3	<i>branch 14</i> : <i>attr_11</i> > 0.197 and <i>attr_27</i> <= 0.8145 and <i>attr_54</i> <= 0.0205 and <i>attr_53</i> > 0.0166: M	1	12	8.33%

On completion of the tree branch ranking and exploration process, the highest ranked branch can potentially be considered as the weakest tree branch because it has the highest error rate of all the branches. It is acknowledged that this is only a speculative

consideration and ranking on the error rate of the branches is indeed a simplistic comparison and does not take into account the overall proportion of the samples involved. These issues are addressed in Session 4.4 as part of the experiment analysis and discussion.

When checking the contents of the three weakest tree branches ranked in the top three positions, the most error-sensitive attribute *attr\_11* can be seen in all these three weakest branches. However, this raises some questions about the absence of the other three highly ranked attributes from these potentially weakest tree branches and it is reasonable to ask why there is no sign of *attr\_48* which ranked second, *attr\_45* which ranked third, and *attr\_9* which ranked fourth in any of the top three ranked weakest tree branches with various value ranges, as shown in the lower sub-table in Table 4-1.

One way to verify the potentially most error-sensitive attributes identified in the process is to filter out various combinations of these attributes from the re-classification process to see if the classification performance can be improved, and though the results are mixed, they are encouraging. The three test cases are shown in Table 4-2.

The first test case shows that after filtering out the most error-sensitive attribute *attr\_11*, the re-classification accuracy decreased from 71.15% to 69.23% and four new errors were added. On the other hand, the second test case shows that the accuracy rate increased from 71.15% to 72.60% after filtering out only the second attribute *attr\_45*, and the number of errors reduced from 60 to 57, and the third test case shows that the accuracy rate increased from 71.15% to 72.12% after filtering out the second and third most error-sensitive attributes, *attr\_45* and *attr\_48*, with the number of errors reducing from 60 to 58.

Table 4-2 - Pre- vs. Post- removal of error-sensitive attributes in the Sonar dataset

Original run of J48/C4.5 classification with all attributes	Re-run J48/C4.5 without error-sensitive attribute attr_11
=== Stratified cross-validation === === Summary === Correctly Classified Instances 148 <b>71.15%</b> Incorrectly Classified Instances 60 28.85 %	=== Stratified cross-validation === === Summary === Correctly Classified Instances 144 <b>69.23%</b> Incorrectly Classified Instances 64 30.77%
Re-run J48/C4.5 without error-sensitive attribute attr_45	Re-run J48/C4.5 without error-sensitive attribute attr_45 and attr_48
=== Stratified cross-validation === === Summary === Correctly Classified Instances 151 <b>72.60%</b> Incorrectly Classified Instances 57 27.40%	=== Stratified cross-validation === === Summary === Correctly Classified Instances 150 <b>72.12%</b> Incorrectly Classified Instances 58 27.88%

#### 4.3.2 The MAGIC Gamma Telescope Dataset

This relatively large binary dataset was generated by a Monte Carlo program to simulate the registration of gamma particles from the atmospheric Cherenkov gamma telescope to study the evolution of electromagnetic showers initiated by the gammas and other particles in the atmosphere. It has 19,020 samples and ten attributes which represent several characteristic parameters of electromagnetic particles, such as axis of the ellipse for elongated clusters [HKC98].

The initial classification process correctly classified 11,383 samples with an accuracy rate of 85.06%, and 2,842 samples were misclassified. The resulting decision tree has 629 nodes and 315 branches, as shown in Figure 4-4.

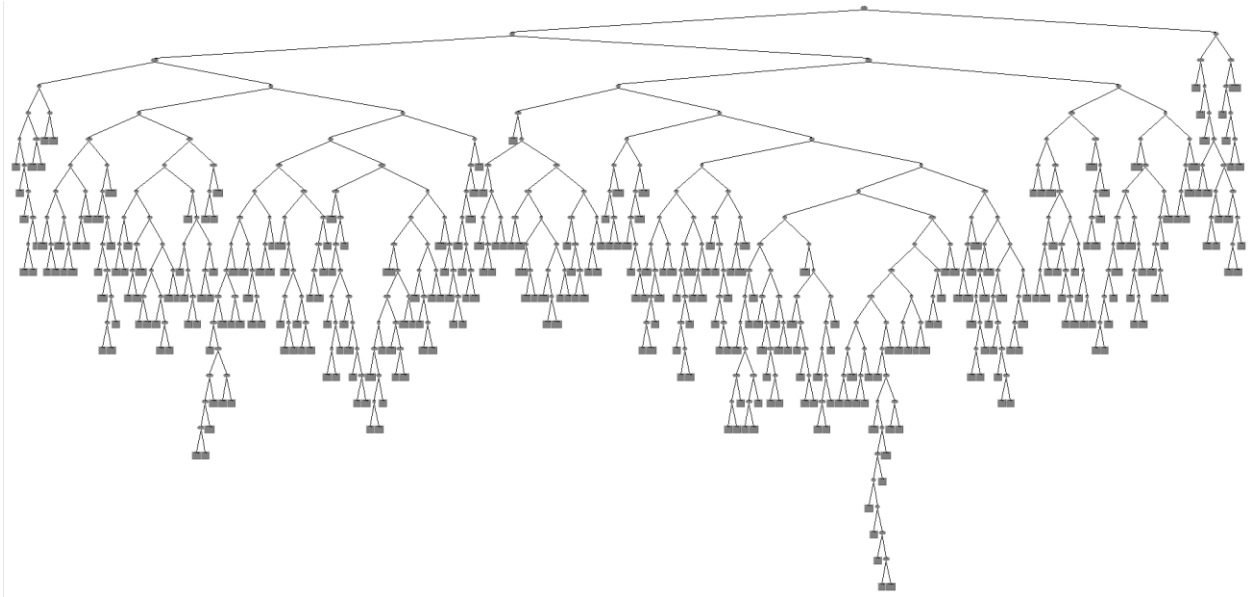


Figure 4.4 - Decision tree from the MAGIC Gamma Telescope dataset

On completion of the error-sensitive attribution evaluation process, attributes *fAlpha*, *fAsym*, *fM3Long* and *fWidth* are ranked as potentially the four most error-sensitive parameters. When comparing the ranking list generated by the gain ratio algorithm and the information gain algorithm, the top ranked attribute *fAlpha* is also ranked as the most significant parameter by information gain but is only ranked second by gain ratio. The second ranked *fAsym* is ranked fourth by gain ratio and is not in the top four by information gain. The fourth ranked *fWidth* is ranked second by information gain but third by gain ratio, as shown in the upper sub-table in Table 4-3.

On completion of the tree branch ranking and exploration process and after checking the contents of the three weakest tree branches ranked in the top three positions, the top and fourth most error-sensitive attributes, *fAlpha* and *fWidth*, can be seen in all three weakest branches with various value ranges, as shown in the lower sub-table in Table 4-3.



Table 4-3 - Exploration of error-sensitive branches in the Gamma dataset

Rank	Attribute-error counter	Gain ratio	Info gain		
1	fAlpha (865): 18.78~43.99	fLength (0.10211)	fAlpha (0.1771)		
2	fAsym (763): -18.29~3.27	fAlpha (0.06127)	fWidth (0.1324)		
3	fM3Long (746): -2.85~17.81	fWidth (0.05037)	fLength (0.1158)		
4	fWidth (614): 18.59~28.80	fAsym (0.0482)	fM3Long (0.1044)		

Rank	Root-to-leaf decision path (classification rule)	Error count	Sample count	Error rate
1	<b>branch 361</b> : fLength <= 114.58 and <b>fAlpha</b> > 20.25 and fLength <= 38.54 and fSize > 2.32 and <b>fWidth</b> > 9.9 and <b>fAlpha</b> > 39.64 and fDist > 165.71 and fLength <= 27.41 and fSize > 2.44 and <b>fWidth</b> > 14.29 and fConc1 <= 0.28: g	63	129	48.84%
2	<b>branch 384</b> : fLength <= 114.58 and <b>fAlpha</b> > 20.25 and fLength > 38.54 and <b>fAlpha</b> <= 31.89 and fLength <= 63.77 and <b>fWidth</b> > 11.72 and fLength > 43.1 and <b>fM3Long</b> <= 30.7 and fConc <= 0.41 and <b>fWidth</b> <= 29.68 and fDist > 98.29: g	27	59	45.76%
3	<b>branch 290</b> : fLength <= 114.58 and <b>fAlpha</b> > 20.25 and fLength <= 38.54 and fSize > 2.32 and <b>fWidth</b> > 9.9 and <b>fAlpha</b> <= 39.64 and fConc > 0.55 and fSize > 2.4 and fSize <= 2.57 and fConc1 <= 0.45: g	33	75	44%

To verify the potentially most error-sensitive attributes identified in the process, various combinations of these highly error-sensitive attributes are filtered out from the re-classification process. The results show that when only the most error-sensitive attribute *fAlpha* is filtered out, accuracy decreases from 85.06% to 80.59%, and when the two most error-sensitive attributes, *fAlpha* and *fAsym*, are filtered out, accuracy decreases to 80.75%. However, when only the second ranked most error-sensitive attribute *fAsym* is filtered out, accuracy increases to 85.17%, which is 22 less errors compared with the initial result, as shown in Table 4-4.

Table 4-4 - Pre- vs. Post- removal of error-sensitive attributes in the Gamma dataset

Original run of J48/C4.5 classification with all attributes	Re-run J48/C4.5 without error-sensitive attributes fAlpha
=== Stratified cross-validation ===	=== Stratified cross-validation ===
=== Summary ===	=== Summary ===
Correctly Classified Instances 16,178 <b>85.06%</b>	Correctly Classified Instances 15,329 <b>80.59%</b>
Incorrectly Classified Instances 2,842 14.94%	Incorrectly Classified Instances 3,691 19.41%
Re-run J48/C4.5 without error-sensitive attributes fAlpha and fAsym	Re-run J48/C4.5 without error-sensitive attributes fAsym
=== Stratified cross-validation ===	=== Stratified cross-validation ===
=== Summary ===	=== Summary ===
Correctly Classified Instances 15,359 <b>80.75%</b>	Correctly Classified Instances 16,200 <b>85.17%</b>
Incorrectly Classified Instances 3,661 19.25%	Incorrectly Classified Instances 2,820 14.83%

### 4.3.3 The Yeast Dataset

This dataset in its original form was used to predict Cellular Localization Sites of Proteins in gram-negative bacteria, but it has been converted to a binary dataset for the testing purposes of this study. It has 1,484 samples and eight meaningful attributes [NK91] [HN96].

The initial classification process correctly classified 1,116 samples with an accuracy rate of 75.20%, and 368 samples were misclassified. The resulting decision tree has 89 nodes and 45 branches, as shown in Figure 4.5.

Figure 4.5 - Decision tree from the Yeast dataset

Meanwhile, after the completion of the tree branch ranking and exploration process and after checking the contents of the three weakest tree branches, all top four most error-sensitive attributes, *nuc*, *gvh*, *alm* and *mcg*, can be seen in all three weakest branches with various value ranges, as shown in the lower sub-table in Table 4-5.

Table 4-5 – Exploration of error-sensitive branches in the Yeast dataset

Rank	Attribute-error counter	Gain ratio	Info gain	
1	nuc (111): 0.25~0.33	nuc (0.062)	nuc (0.096)	
2	gvh (102): 0.46~0.52	pox (0.061)	alm (0.064)	
3	alm (92): 0.49~0.53	alm (0.042)	mcg (0.058)	
4	mcg (78): 0.45~0.52	mit (0.028)	gvh (0.037)	

Rank	Root-to-leaf decision path (classification rule)	Error count	Sample count	Error rate
1	<i>branch 13</i> : <u>alm</u> > 0.4 and <u>nuc</u> > 0.24 and mit <= 0.4 and <u>mcg</u> <= 0.58 and <u>nuc</u> <= 0.31 and <u>alm</u> > 0.52: POS	42	105	40.00%
2	<i>branch 12</i> : <u>alm</u> > 0.4 and <u>nuc</u> > 0.24 and mit <= 0.4 and <u>mcg</u> <= 0.58 and <u>nuc</u> <= 0.31 and <u>alm</u> <= 0.52 and vac > 0.46: NEG	30	88	34.09%
3	<i>branch 8</i> : <u>alm</u> > 0.4 and <u>nuc</u> <= 0.24 and <u>nuc</u> > 0.14 and <u>gvh</u> > 0.39: NEG	91	562	16.19%

Various combinations of the most error-sensitive attributes are filtered out in re-runs of the classification process to check for a possible direct association between the most error-sensitive attributes and the classification results. The results show that when only the top most error-sensitive attribute *nuc* is filtered, accuracy decreases from 75.20 % to 71.56%, most likely because *nuc* is also the most significant attribute as ranked by gain ratio used in the J48/C4.5 decision tree classifier. On the other hand, when the second and fourth most error-sensitive attributes, *gvh* and *mcg*, are filtered out for a re-run of classification, accuracy increases to 76.01% with a reduction of 12 errors. When adding the third most error-sensitive attribute, *alm*, to the filtering list, the accuracy rate decreases again to 73.99%, as shown in Table 4-6.

Table 4-6 - Pre- vs. Post- removal of error-sensitive attributes in the Yeast dataset

Original run of J48/C4.5 classification with all attributes	Re-run J48/C4.5 without the most error-sensitive attribute nuc
=== Stratified cross-validation ===	=== Stratified cross-validation ===
=== Summary ===	=== Summary ===
Correctly Classified Instances 1,116 <b>75.20%</b>	Correctly Classified Instances 1,062 <b>71.56%</b>
Incorrectly Classified Instances 368 24.80%	Incorrectly Classified Instances 42 23.99%

Re-run J48/C4.5 without error-sensitive attributes gvh & mcg	Re-run J48/C4.5 without error-sensitive attributes mcg, gvh and alm
=== Stratified cross-validation ===	=== Stratified cross-validation ===
=== Summary ===	=== Summary ===
Correctly Classified Instances 1,128 <b>76.01%</b>	Correctly Classified Instances 1,098 <b>73.99%</b>
Incorrectly Classified Instances 356 23.99%	Incorrectly Classified Instances 386 26.01%

Despite the uninspiring 1.21% classification accuracy reduction when three highly ranked error-sensitive attributes are filtered out, the positive finding is that the resulting decision tree structure is considerably simplified and becomes a tree with only seven nodes and four branches, as shown in Figure 4.6. Section 4.4 discusses the possible benefits from this decision simplification trade off over complexity with marginal enhancement.

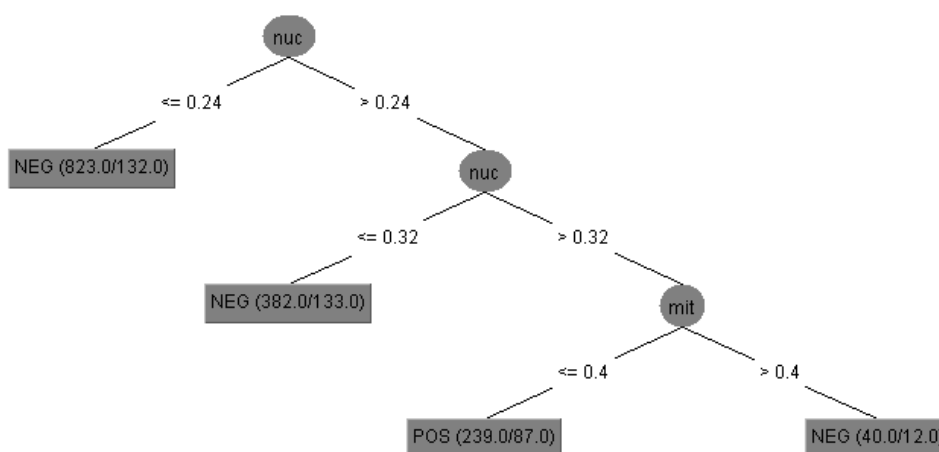


Figure 4.6 – Greatly simplified decision tree after removing highly error-sensitive attributes

#### 4.3.4 The Cardiocotography Dataset

This is a dataset of diagnostic attribute values recorded on cardiocotograms to measure various attributes of fetal heart rate (FHR) and uterine contraction (UC). It contains 2,126 samples and 21 meaningful attributes. The original three class labels are regrouped into two groups to convert the original data into a binary dataset for the testing purpose of this study [ABG00].

The initial classification process correctly classifies 1,981 samples with an accuracy rate of 93.18% and 145 samples are misclassified. The resulting decision tree has 95 nodes and 48 branches, as shown in Figure 4.7.

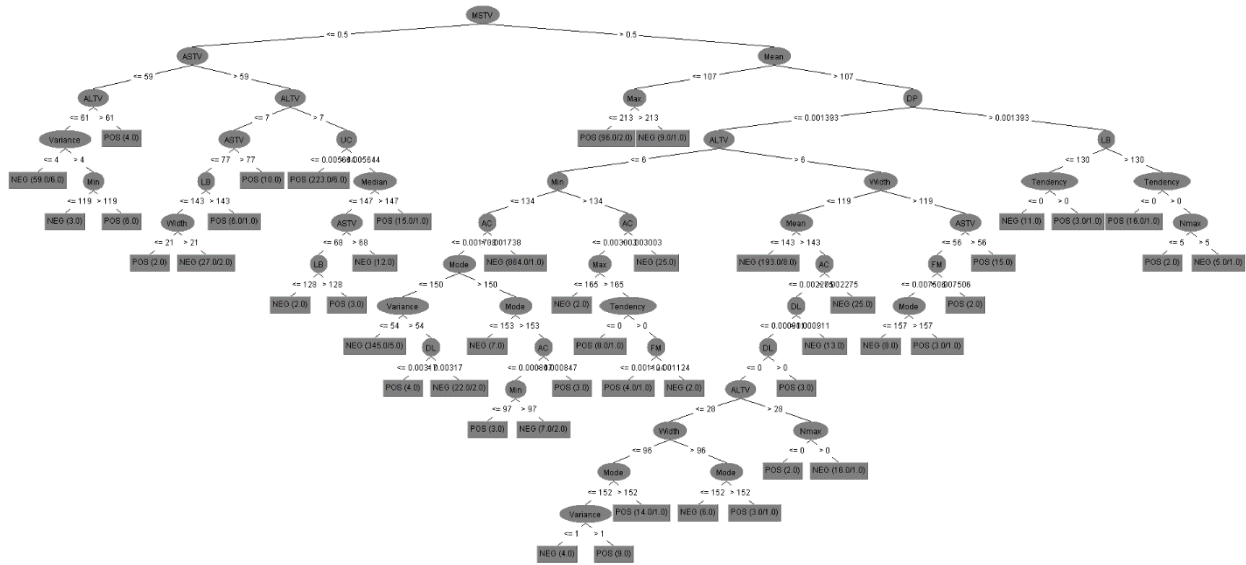


Figure 4.7 - Decision tree from the Cardiocotography dataset

On completion of the error-sensitive attribution evaluation process, attributes *ASTV*, *ALTV*, *AC* and *MLTV* are ranked as potentially the four most error-sensitive attributes. When comparing the ranking list generated by the gain ratio algorithm and the information gain algorithm, the top ranked attribute *ASTV* is also ranked as the most significant

attribute by information gain but is ranked fourth by the gain ratio. The other three highly ranked error-sensitive attributes, *ALTV*, *AC* and *MLTV*, on the other hand, are not included in the top four ranked significant attributes by the gain ratio, and only *ALTV* and *AC* are included in the top four ranked significant attributes by information gain, as shown in the upper sub-table in Table 4-7.

On completion of the tree branch ranking and exploration process and checking the contents of the three weakest tree branches, the top three most error-sensitive attributes, *ASTV*, *ALTV*, *AC*, can be seen in all three weakest branches with various value ranges. However, the fourth ranked attribute *MLTV* is absent in all tree branches, just as it is absent in all highly ranked significant attributes either by gain ratio or information gain, as shown in the lower sub-table in Table 4-7.

Table 4-7 - Exploration of error-sensitive branches in the Cardiotocography dataset

Rank	Attribute-error counter	Gain ratio	Info gain	
1	ASTV (90): 42.47~ 62.89	DP (0.2322)	ASTV (0.23206)	
2	ALTV (46): 5.04~ 26.72	DS (0.2201)	MSTV (0.22895)	
3	AC (43): 3.13~3.98	MSTV (0.1323)	AC (0.20366)	
4	MLTV (38): 6.37~ 8.71	ASTV (0.1267)	ALTV (0.15595)	

Rank	Root-to-leaf decision path (classification rule)	Error count	Sample count	Error rate
1	<i>branch 1</i> : MSTV <= 0.5 and <u>ASTV</u> <= 59 and <u>ALTV</u> <= 61 and Variance <= 4: NEG	6	59	10.17%
2	<i>branch 18</i> : MSTV > 0.5 and Mean > 107 and DP <= 0.001393 and <u>ALTV</u> <= 6 and Min <= 134 and <u>AC</u> <= 0.001738 and Mode <= 150 and Variance > 54 and DL > 0.00317: NEG	2	22	9.09%
3	<i>branch 29</i> : MSTV > 0.5 and Mean > 107 and DP <= 0.001393 and <u>ALTV</u> > 6 and Width <= 119 and Mean <= 143: NEG	8	193	4.15%

To verify the potentially most error-sensitive attributes identified in the process, various combinations of these attributes are filtered out from the re-classification process. The results show that when only the most error-sensitive attribute *ASTV* is filtered out, accuracy decreases from 93.18% to 92.10%, but when the other three highly ranked error-sensitive attributes, *ALTV*, *AC* and *MLTV*, are filtered out, accuracy increases to 93.27%; and when only the fourth ranked most error-sensitive attribute *MLTV* is filtered out, accuracy increases further to 93.46%, even though *MLTV* is not considered to be a highly significant attribute and is also absent from the resulting decision tree structure, as shown in Table 4-8.

Table 4-8 - Pre- vs. Post-removal of error-sensitive attributes in the Cardiotocography dataset

Original run of J48/C4.5 classification with all attributes	Re-run J48/C4.5 without error-sensitive attribute ASTV
==== Stratified cross-validation ====	==== Stratified cross-validation ====
==== Summary ====	==== Summary ====
Correctly Classified Instances 1,981 <b>93.18%</b>	Correctly Classified Instances 1,958 <b>92.10%</b>
Incorrectly Classified Instances 145 6.82%	Incorrectly Classified Instances 168 7.90%
Re-run J48/C4.5 without error-sensitive attributes ALTV, AC and MLTV	Re-run J48/C4.5 without error-sensitive attribute MLTV
==== Stratified cross-validation ====	==== Stratified cross-validation ====
==== Summary ====	==== Summary ====
Correctly Classified Instances 1,983 <b>93.27%</b>	Correctly Classified Instances 1,987 <b>93.46%</b>
Incorrectly Classified Instances 143 6.73%	Incorrectly Classified Instances 139 1.08%

#### 4.3.5 The Glass Identification Dataset

This is a forensic science dataset containing oxide content and property information on window glass samples to determine their specific types of glass. It has 214 samples and



nine attributes for eight oxide contents and one property of reflectiveness for the window samples. The original dataset has seven class values which are converted into two labels, POS for windows that were “float” processed, and NEG for all other types.

The initial classification process correctly classified 175 samples with an accuracy rate of 81.78%, and 39 samples were misclassified. The resulting decision tree has 33 nodes and 17 branches, as shown in Figure 4.8.

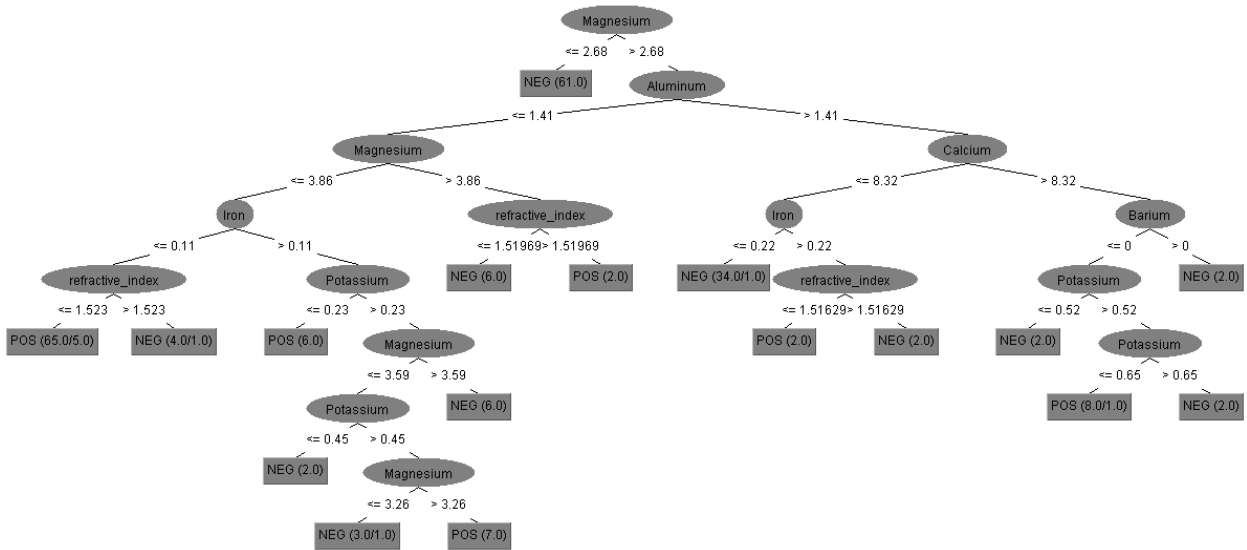


Figure 4.8 - Decision tree from the Glass ID dataset

On completion of the error-sensitive attribution evaluation process, attributes *aluminum*, *magnesium*, *sodium* and *calcium* are ranked as potentially the four most error-sensitive attributes. When comparing the ranking list generated by the gain ratio algorithm and the information gain algorithm, the top ranked attribute *aluminum* is ranked as the second most significant attribute by both the gain ratio and information gain. On the other hand, the second ranked attribute *magnesium* is ranked as the top most significant attribute by both the gain ratio and information gain. The third ranked attribute *sodium* is not ranked as

a highly significant attribute by either the gain ratio or information gain, and the fourth ranked attribute *calcium* is also ranked fourth by information gain but is not ranked as highly significant by the gain ratio, as shown (or missing) in the upper gain ratio and information gain ranking sub-tables in Table 4-9.

Table 4-9 - Exploration of error-sensitive branches in the Glass ID dataset

Rank	Attribute-error counter	Gain ratio	Info gain	
1	Aluminum (27): 1.17~1.63	Magnesium (0.312)	Magnesium (0.2694)	
2	Magnesium (24): 2.09~3.55	Aluminum (0.265)	Aluminum (0.2643)	
3	Sodium (8): 13.28~13.50	Barium (0.154)	refractive_index (0.2173)	
4	Calcium (6): 8.79~9.07	refractive_index (0.142)	Calcium (0.171)	

Rank	Root-to-leaf decision path (classification rule)	Error count	Sample count	Error rate
1	<i>branch 6</i> : <b>Magnesium</b> > 2.68 and <b>Aluminum</b> <= 1.41 and <b>Magnesium</b> <= 3.86 and Iron > 0.11 and Potassium > 0.23 and <b>Magnesium</b> <= 3.59 and Potassium > 0.45 and <b>Magnesium</b> <= 3.26: NEG	1	3	33.33%
2	<i>branch 3</i> : <b>Magnesium</b> > 2.68 and <b>Aluminum</b> <= 1.41 and <b>Magnesium</b> <= 3.86 and Iron <= 0.11 and refractive_index > 1.523: NEG	1	4	25.00%
3	<i>branch 2</i> : <b>Magnesium</b> > 2.68 and <b>Aluminum</b> <= 1.41 and <b>Magnesium</b> <= 3.86 and Iron <= 0.11 and refractive_index <= 1.523: POS	5	65	7.69%

On completion of the tree branch ranking and exploration process and after checking the contents of the three weakest tree branches, two of the three most error-sensitive attributes, *aluminum* and *magnesium*, can be seen in all three weakest branches with various value ranges. However, the third ranked attribute *sodium* is absent in all tree branches, just as it is absent in all highly ranked significant attributes either by gain ratio or information gain, as shown in the lower sub-table in Table 4-9, but the fourth ranked attribute

*calcium* can be seen in some limited tree branches, as shown in the lower sub-table in Table 4-9.

To verify the potentially most error-sensitive attributes identified in the process, various combinations of these attributes are filtered out from the re-classification process. The results show that when the most error-sensitive attribute *aluminum* is filtered out, accuracy increases from 81.78% to 84.11%, and when the third most error-sensitive attribute *sodium* is filtered out, accuracy increases to 82.71%, and when both the third and fourth most error-sensitive attributes *sodium* and *calcium* are filtered out, accuracy increases to 84.58%, as shown in Table 4-10.

Table 4-10 - Pre- vs. Post-removal of error-sensitive attributes in the Glass ID dataset

Original run of J48/C4.5 classification with all attributes			Re-run J48/C4.5 without error-sensitive attributes Aluminum		
=== Stratified cross-validation ===			=== Stratified cross-validation ===		
=== Summary ===			=== Summary ===		
Correctly Classified Instances	175	81.78%	Correctly Classified Instances	180	84.11%
Incorrectly Classified Instances	39	18.22 %	Incorrectly Classified Instances	34	15.89%
Re-run J48/C4.5 without error-sensitive attributes Sodium			Re-run J48/C4.5 without error-sensitive attributes Sodium & Calcium		
=== Stratified cross-validation ===			=== Stratified cross-validation ===		
=== Summary ===			=== Summary ===		
Correctly Classified Instances	177	82.71%	Correctly Classified Instances	181	84.58%
Incorrectly Classified Instances	37	17.29 %	Incorrectly Classified Instances	33	15.42%

#### 4.4 EXPERIMENT ANALYSIS

The exploration study on the weakest and most error-sensitive branches of a decision tree can be considered as a natural progression from the evaluation study of error-sensitive

attributes in a classification task, and it is intuitive to assume that the weakest branches would contain the most error-sensitive attributes in their nodes as a form of close relationship – causation or correlation, or a bit of both. The experiments on five UCI datasets illustrate such a close relationship and have highlighted some interesting issues.

#### **4.4.1 Comparison Between the Gain Ratio and Information Gain Approach**

In this study, the criterion of determining the weakest and most error-sensitive tree branches is the error rate, which means, the higher the ratio of the number of classification errors over the number of samples on a tree branch, the higher the risk of having a classification error on this branch therefore making it a weaker branch. An alternative criterion can be to use the actual error count, so the higher the error value of a tree branch, the higher the risk of this branch, therefore making it a weaker branch.

If the error rate approach is compared to the gain ratio approach [Qui86], then the error count approach can also be compared to the information gain approach [Sha48]. While the error count approach may be problematic in relation to bias in favour of a branch with a higher number of samples, it is a clear indicator of the level of significance of a branch in terms of the number of samples involved. On the other hand, the error rate approach may help overcome this bias issue as its solution of using a ratio can be problematic as some significant branches may be missed out but with lower error counts in the training data.

Take one example from the Gamma dataset. The initial classification process generates a decision tree with 629 nodes and 315 branches, as shown in Figure 4.4. During the branch ranking and exploration process, it is revealed that branch No.4 has one error out

of three samples on this branch so its error rate is 33.33%, branch No.133 has 18 errors out of 96 samples on this branch so its error rate is 18.75%, and branch No.62 has 130 errors out of 3,950 samples on this branch so its error rate is 3.29%, as shown in Table 4-11.

When using the error rate criteria, branch No.4 with an error rate of 33.33% is ranked higher than branch No.62 with error rate of 3.29%; when using the error count criteria, branch No.62 with 130 errors out of 3,950 samples is ranked higher than branch No.4 with one error from three samples. It is reasonable to think that branch No.62 with 130 errors out of 3,950 samples is more significant than branch No.4 in this instance because of the significant number of samples involved, but it is also logical to doubt its level of risk because the error rate of 3.29% is rather low compared to other highly ranked attributes all with double-digit error rates.

Table 4-11 – Comparing error rate and error count for three branches in the Gamma dataset

Branch No.	Root-to-leaf decision tree branch (classification rule)	Sample count	Error count	Error rate
4	fLength <= 114.586 and fAlpha <= 20.2535 and fM3Long <= -67.699 and fWidth <= 37.3294 and fAlpha <= 8.2049 and fDist > 159.558 and fAsym > -99.6169 and fAlpha > 3.726 and fLength <= 101.2776 and fDist <= 203.91: NEG	3	1	33.33%
133	fLength <= 114.586 and fAlpha > 20.2535 and fLength <= 38.5309 and fSize <= 2.3304 and fLength > 12.0413 and fWidth > 6.6496 and fLength <= 29.2488 and fM3Long > -19.5781 and fDist <= 217.7938 and fSize > 2.3133 and fConc <= 0.8096: POS	96	18	18.75%
62	fLength <= 114.586 and fAlpha <= 20.2535 and fM3Long > -67.699 and fWidth > 12.1606 and fWidth <= 46.4167 and fAlpha <= 10.051 and fConc1 <= 0.2396 and fWidth <= 31.4988 and fLength <= 99.1202 and fDist > 124.4584: POS	3,950	130	3.29%

A third approach may be to conduct experiments using both approaches, to compare and evaluate their results carefully and subsequently allocate the ranking in a more

balanced way. This kind of balanced approach may require more time and resources than a systematic approach with parameters for self-tuning, but it is more about constructing the context to gain a better understanding of the data rather than just focusing on evolving algorithms to gain better performance on efficiency.

Hypothetically, each attribute has its own characteristics and so does each dataset with many different attributes, therefore, as attributes and datasets change, so should its specific method of evaluation and analysis. So, the use of this error rate approach is just a start, and it is intended as an initial model to open up the exploration and discussion.

Another consideration when adopting the error rate approach is that the gain ratio method is the underlying algorithm for the J48/C4.5 decision tree model used in the experiments, so in an attempt to keep some consistency between the overall classification model and the attribute evaluation model in an early stage of the study, the error rate approach is used in this current study for the experiments.

#### **4.4.2 Effectiveness Considerations**

In terms of the effectiveness of the identification of the weakest branches of a decision tree, the experiment results show supportive signs as reported in earlier section Experiments and in highlighted examples below, but in order to examine their real values and real meaning, the participation of domain experts is essential.

For example, 45 tree branches are generated from the initial classification of the Yeast dataset, and 368 errors from the 1,484 samples. On completion of the weakest branch exploration process, the three weakest branches, branch No.13, branch No.12 and branch No.8, account for 163 errors from 755 samples, as shown in Table 4-5. That is, 6.67%

(3/45) of the branches account for 44.29% (163/368) of the errors from 50.88% (755/1,484) of the samples compared to the overall error rate of 23.99%. While these weakest branches may look significant in terms of error ratio and result comparison, specific biology and chemistry knowledge on protein and molecule analysis will be required to further examine the attributes and value ranges identified from these weakest to help interpret the association and real meaning of their association in order to gain further understanding.

Another example is the Cardiotocography dataset, where 48 branches are generated from the initial classification, and 145 errors from the 2,126 samples. On completion of the weakest branch exploration process, the three weakest branches, branch No.1, branch No.18 and branch No.29, account for 16 errors from 274 samples, as shown in Table 4-7. That is, 6.25% of the branches account for 11.00% (16/145) of errors from 12.89% (274/2126) of samples compared to the overall error rate of 6.73%.

Another benefit of weakest branch exploration is its identification of a specific attribute together with its split point value at each node along each tree branch, so it adds valuable information to the identified error-sensitive attributes and provides further context about the weakest and error-sensitive branch exploration in relation to particular error-sensitive attributes and their specific value ranges.

For example, on completion of the error-sensitive attribute evaluation and weakest branch exploration process for the Glass ID dataset, it identifies that *aluminum* is potentially the most error-sensitive attribute with an ambiguous value range of 1.17~1.63; and *magnesium* is the second most error-sensitive attribute with an ambiguous value range of 2.09~3.55. An examination of the error-sensitive branches shows that branch No.6 is the

weakest branch as it contains the top two most error-sensitive attributes, *aluminum* with its aggregated split point value range as  $\leq 1.41$ , which is close to the top end of its ambiguous value range of 1.17~1.63, and *magnesium* with its aggregated split point value ranges of 2.68~3.26, which is also close to the top end of its ambiguous value range of 2.09~3.55, as shown in Table 4-9.

As for the Yeast dataset, it has identified that attribute *nuc* is potentially the most error-sensitive with an ambiguous value range of 0.25~0.33, *alm* is the third most error-sensitive with an ambiguous value range of 0.49~0.53, and *mcb* is the fourth most error-sensitive with an ambiguous value range as 0.45~0.52. An examination of the error-sensitive branches shows that branch No.13 is the weakest branch as it contains two of the top three most error-sensitive attributes; *nuc* with its aggregated split point value range of 0.24~0.31, is a close match with its ambiguous value range of 0.25~0.33; and *alm* with its aggregated split point value range of  $> 0.52$  is also close to the top end of its ambiguous value range of 0.49~0.53, as shown in Table 4-5.

These experiments and results point to a certain correlation between the weakest branches and the most error-sensitive attributes, and also some connection between certain split point value ranges of the weakest branches and the ambiguous value ranges of the most error-sensitive attributes.

Much work is still required to ensure the current tree branch exploration process is more sophisticated and specific, hence on this note, the importance of domain expertise joining this post-classification analysis must be reiterated. Only the stakeholders of the data and the domain experts can genuinely examine and advise on the correctness and



effectiveness of the proposed findings and can eventually make practical use of the proposed processes. Without close collaboration with the stakeholders and in-depth field knowledge, such a study might be viewed as only playing with numbers and building “toy” projects for no real application.

## **4.5 Related Work**

Decision trees have a reputation for being efficient and illustrative in data mining and classification. They can outline details on decision paths with specific attributes and value ranges, therefore they have been adopted as a classification platform upon which to add and build the weakest branch exploration model for data and error analysis in this study.

### **4.5.1 Decision Tree Classification**

Research on the decision tree model probably started in the 1950s when academics in Europe and America enjoyed a more peaceful and inspirational post-war life, which brought forth a flourishing interest in the study of human thinking and learning behavior, and created a fertile environment to cultivate the various ideas of decision trees and their association with inductive learning and decision making. These early studies were also assisted by the utilization of early computer concepts and technology made available during this period [Ban60] [Fei59] [Hun62] [LR57]. As the development of inductive decision trees advanced further in the 1970s and 1980s, the algorithms and implementation involved became more complex and sophisticated, and some were applied to practical and commercial applications with satisfactory outcomes [BFO84] [Qui86] [Sam93].

With increasing attention and success, the decision tree model started to emerge from its infancy and became to be regarded as a sub-topic of a mainstream subject, such as psychology or mathematics and was gradually viewed as a full-blown research area in the field of machine learning and data mining. Since the 1990s, computer technology has been advancing rapidly every year, generating and collecting large amounts of data on all aspects of life which opens the door for large-scale data mining development and the utilization of decision tree models to achieve efficient and effective learning results. New approaches and enhancements to existing algorithms have been developed to prune and refine decision trees [Dom99] [LYW04] [PD00] [Qui93] [WF05], to make the model more efficient and effective in providing accurate and meaningful results when it is used to explore and analyze a large amount of complex data. The research and enhancement work on decision trees is continuing, and the research itself is also branching out into more specific and focused areas, and each branch tends to focus on a specific factor, for example, the scalability factor, the cost factor and the risk factor.

#### **4.5.2 Tree- and Branch-based Study and Development**

Based on the decision tree platform, the study reported in this chapter focuses on developing an exploration process to identify the weakest branches of a decision tree, to help evaluate error-sensitive attributes and value range patterns as a way to gain further understanding about the correlation between classification errors and value patterns and the dataset as a whole.

This approach to adapting a decision tree as the basis for further development in data analysis is not an isolated one. For example, to improve the accuracy and clarity of

decision trees, various pruning techniques have been developed. Three pruning methods were summarized in one study by Quinlan as early as 1987. The first one is the cost-complexity pruning method, which was initially created in the CART decision tree study [BFO84], and this pruning process requires two stages. In the first stage, a sequence of decision trees is created, starting from the original decision tree and each subsequent tree is created by substituting one or more subtrees with leaves until the final tree becomes just a leaf. In the second stage, the created trees are evaluated and the one with the best score from its cost-complexity function is selected as the optimal model.

The second method is the reduced error method. This pruning process starts from leaf level and replaces the upper node with the leaf carrying the majority class label, then checks the overall accuracy and if the error rate has been reduced or remains the same, it keeps the replacement, however if not, then the replacement is not kept. This replacement and test routine goes on to the next node until all nodes are covered and the last resulting tree is considered the optimal model.

The third method is the pessimistic method. It uses the concept of “continuity correction” in statistics to estimate the error rate of each node and compares it with the standard error assuming errors are binomially distributed. If a subtree contains a leaf and the estimated error rate of this leaf is not greater than one standard error, then the subtree is pruned and replaced by this leaf.

Other pruning methods have also been developed over the years, such as the critical value pruning method which creates a critical value threshold based on the value of information gain measured during node construction, so if the information gain of a node is

smaller than the estimated critical value threshold, then the node is pruned and replaced by a leaf [Min89]; and the Minimum Description Length (MDL) pruning methods that are pruning and selecting an optimal tree in various ways but are based on the maximal compression rate of the data [QR89] [MRA95].

A more recently developed approach includes significance tests in the pruning algorithms to identify the superfluous and “just by chance” parts of tree nodes to prune and the genuine and relevant parts of tree nodes to keep to help overcome the typical “overfitting” problem in decision tree classification with improved comprehensibility on tree structure and enhanced accuracy [Fra00].

While the studies and developments on decision tree pruning have been extensive, much of their focus is on classification accuracy enhancement and tree size reduction to eliminate as many errors as possible and very few suggestions on error patterns and error-sensitive value patterns have been collected and analyzed as a part of the study in the context of data analysis.

This is one significant point of difference between this study and the vast amount of research work mentioned earlier. One possible explanation for this is, the measure of progress and success in performance is simple and universal and can be compared by accuracy and error rate in easy and numerical terms, or, in terms of processing time and system resources consumed. However, as for the identification and analysis of error-sensitive attributes and value patterns, the measure of progress and success may not be so obvious. Each dataset has its own set of attributes, and each attribute has its own characteristics, so there is no simple and universal measure to compare the different error and value

patterns between different datasets, and specific domain knowledge and opinions from field experts may be required in order to review and validate the findings.

#### **4.6 Summary**

Every tree in a forest is different and every branch on a tree is different. Each branch may be different in size and shape and also may be different in strength and in the formation of its nodes, leaves and buds. Due to internal factors such as weight and strength, and external factors such as pests and diseases, each branch may also have a different sustainability level in maintaining health and a different risk level in relation to suffering infection and damage. Though some branches may be healthier and stronger than the others, some branches may be considered weaker and more prone to infection and damage.

Likewise, each branch in a decision tree is different, therefore each underlying decision path is different, and so is the error rate resulting from each decision path and its representative branch.

This chapter introduced a tree branch exploration process that explores and identifies the most error-sensitive decision tree branches in terms of attributes and value ranges in a systematic way, and with cross-reference to the error-sensitive attribute evaluation model to highlight the specific attribute and value patterns involved, to examine the possible correlation between errors and such value patterns as a way to gain a better understanding of the data and to raise the stakeholders' awareness of these specific error-sensitive patterns.

Although this weakest branch exploration process is a natural progression from the earlier error-sensitive attribute evaluation process, there is plenty of room for improvement.

One improvement attempt is to enumerate and transform the decision tree structure in a digital map to track and store the classification results and statistics for each node and leaf locally and individually. The detail of this decision tree enumeration process is discussed in the next chapter by progressing to a new stage to seek better transparency and clarity in error analysis.

## CHAPTER 5

### **Enumeration of Decision Trees for Pattern Analysis with Transparency**

Errors are inevitable during data classification, and identifying a particular part of the classification model which may be more susceptible to error than others and uncovering specific error-sensitive value patterns can be challenging. Utilizing the progress made in the weakest branch exploration process, this chapter discusses a decision tree enumeration process to map and register classification statistics to node level for easy retrieval and analysis.

Based on the progress made in the earlier stages of this research journey, this new stage narrows the scope of the problem by focusing on decision trees as a pilot model to develop a simple and effective tagging method to digitize individual nodes of a binary decision tree for node-level analysis, to explore a systematic way of identifying the most problematic and error-prone nodes in a decision tree which can be compared to the Achilles' heel of a classification model. It also links and tracks the classification statistics for each node in a transparent way to identify and examine the potentially “weakest” nodes and error-sensitive value patterns in decision trees to assist cause analysis and enhance development.

This digitization method is not an attempt to re-develop or transform the existing decision tree model. Rather, it involves a pragmatic node ID formulation that crafts numeric values to reflect the tree structure and decision making paths to expand post-classification analysis to detailed node-level. Initial experiments have obtained successful results

in locating potentially high-risk attributes and value patterns which is an encouraging sign to indicate that this study is worth further exploration.

### 5.1 Initial Ideas about Tree Enumeration for Error Evaluation

One key objective of this study is to find the most problematic and error-sensitive part of a classification model which requires the collection, identification and comparison of the classification statistics of its individual component parts. Decision trees have been selected as the pilot model for this study because they are a well-researched classification model with a simple structure where decisions on attributes and values are clearly displayed in a form of branches and nodes, as shown in Figure 5.1.

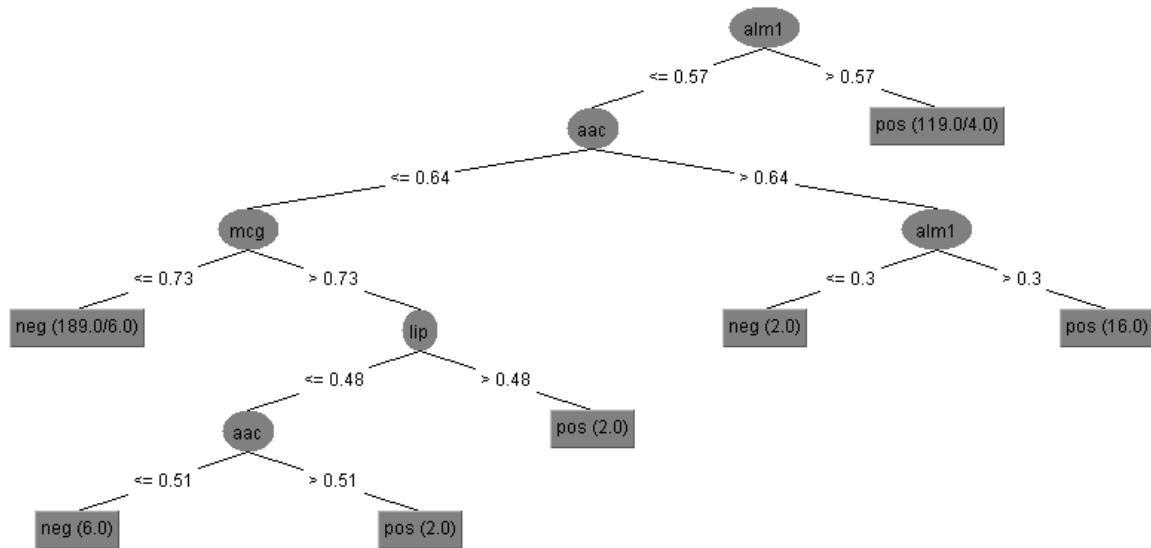


Figure 5.1 - A decision tree example

Using the first branch of the above decision tree as an example:

$alm1 \leq 0.57$  and  $aac \leq 0.64$  and  $mcg \leq 0.73$ : neg (189.0/6.0)

This branch contains three nodes with three split point values, (1)  $\leq 0.57$  for attribute *alm1*, (2)  $\leq 0.64$  for attribute *aac*, and (3)  $\leq 0.73$  for attribute *mcg*. While these



three split points play a role in leading to the six classification errors amongst the 189 instances along this classification path in the form of a decision tree branch, one key question is, which node and its related attribute and value may be more susceptible to the six errors? When expanding this node-specific examination to all branches of the decision tree model, the question can then be generalized as: are some tree nodes more error-prone and error-sensitive than others? If so, can the most error-sensitive nodes and their related attributes and values be identified in a systematic way? These questions are now the focus of this study, and while an earlier report has documented some of the findings [Wu15], more details have now been included for further discussion and analysis.

The rest of this chapter is organized as follows. Section 5.2 describes some initial questions and thoughts that led to the decision tree digitization idea. Section 5.3 outlines the major steps in the decision tree digitization process. Section 5.4 summarizes the experiments on five datasets. Section 5.5 discusses the experiment results and their implications. Section 5.6 reviews some early influential work that inspired this study. Finally, Section 5.7 concludes the current progress and outlines a plan for future exploration.

## **5.2 Background and Assumptions**

Decision trees provide an easy-to-follow graphical view of the classification process, outlining each classification rule from root to leaf step-by-step in the form of node-by-node. If the data volume and the number of attributes of a dataset are not big, or only a stratified portion of samples has been selected for a pilot classification study, then the resulting classification tree structure of a reasonable size can sometimes provide a convenient overview of the classification rules or some key attributes and value ranges involved at a glance.

One issue with such visual representation is, when a large dataset is used and the decision tree structure becomes complex, its graphical view can be clustered and muddled by the full-blown mass of crisscrossing branches and nodes, even after a pruning feature is applied. A complex tree structure can obscure the identification of significant attributes and values when a detailed analysis is required on important classification rules and components.

Another issue is, when node-level statistics are required in a detailed analysis, the visual space reserved for each node on a decision tree may not be the most suitable place to present its node-level statistic values, as this would cluster the presentation of existing branches and nodes even further, making the node scanning and visual interpretation process even more difficult.

One possible solution to address these issues is to provide a unique tag for each tree node, to collect and maintain the node-level classification statistics away from the tree structure and to link them with their respective node by using the unique node tag as the retrieval key. As a result, the classification statistics of each node can be stored and analyzed without any convoluted addition to the existing tree structure.

This decision tree digitization method and error-sensitive pattern analysis may seem trivial; however, if it can be validated as an effective process in tagging and mapping each rule and key components uniquely with individual classification statistics, then it is possible to refine and apply this enumeration and detailed tracking process to other classification models to help illustrate the classification path and step-by-step transition of

statistics in a systematic and transparent way which inspires this study and the digitization development.

### 5.3 Enumeration Process of a Decision Tree

A decision tree model can be transformed into an array of branches where each branch consists of an array of nodes and each node represents its underlying attribute and split point value condition. Because each node can be considered as a child node of its immediate parent node and all levels of parent nodes can be traced back to the root node as the origin, each node can therefore be uniquely identified by a form of regression or inference process based on its hierarchical position within the tree and with the root as the starting point.

Subsequently, a graphical tree can be mapped into a matrix of referential and digitized node IDs in a form of an enumerated map of node IDs, which can link and retrieve node level classification statistics for detailed analysis. The following process summary outlines some key steps of this enumeration process:

**Input:** (1) Initial classification result in the form of a decision tree with **m** branches  
 (2) Original dataset of **n** samples with their newly classified result labels

**Output:** (1) An enumerated map of individual node IDs based on the decision tree  
 (2) A collection of node level classification statistics

**Enumeration process in two stages:**

**Stage-1:** Construct an enumerated map of the resulted decision tree

for 1 to **m** branches of the decision tree from root to leaf

for all nodes in the current branch

(1) if a node is the root node then assign “1” as the starting value of its node ID

(2) if a node is an immediate child node from the root node

then first append a “.” to the current node ID as a node delimiter, and

then add x to form 1.x as its 2nd part of the node ID, x denotes the current number of immediate child nodes branching out from the root and increments by 1 by counting from left to right, e.g. 1.1 as the 1st child node, 1.2 as the 2nd child node, and so on

(3) if a non-root node has child nodes  
     then first append a “.” to the current node ID as a node delimiter, and  
     then assign 1 to its 1st child node on the left as its node ID, assign 2 to its 2nd child node on the right as its node ID; a node ID example is: 1.1.2.2.2.1

**Stage-2:** Traverse and collect individual node level classification statistics

for 1 to **n** data sample  
   for 1 to **m** branches in the enumerated map  
     for all nodes in the current branch  
       (1) if current sample’s attribute value satisfies current node’s split-point condition  
         then continues to next node along current branch  
       (2) if current node’s split point condition cannot be satisfied  
         then advance to the start of next branch in the map  
       (3) if end of current branch is reached and the leaf-node condition is satisfied  
         then update and store the node-level statistics using the node ID as the key for all nodes of the current branch, and  
         then move to the next data sample

On completion of the tree digitization and statistics collection process, a simple ranking of the classification error rate by node IDs can potentially reveal the “weakest” and most error-sensitive node in the tree. The word “potentially” has to be highlighted and emphasized here. Using a node’s error rate value instead of its error count number may avoid the bias towards “heavy traffic” nodes which are associated with higher counts of instances; however, this may unduly magnify the weakness of some “low traffic” nodes which are associated with lower counts of instances. For example, node-A has been traversed by ten samples with five errors so its error rate is 50%, node-B has been traversed by 100 instances with 48 errors so its error rate is 48%. While node-A is subsequently ranked as a weaker node than node-B by comparing the error rate, this may not necessarily be true when more data are used for testing. In recognizing this issue, significant testing and threshold value control on the selection criteria can be implemented as an enhancement measure in a later development.

This decision tree digitization process can be considered as another step forward in the study of error-sensitive value patterns in data classification and the results from the initial experiments appear to support this idea.

## 5.4 Experiments

The five UCI datasets that are used in the experiments for the error-sensitive attribute evaluation process, Ecoli, Pima Indians diabetes, Wisconsin Cancer, Liver Disorders and Page Blocks are tested again in this decision tree enumeration process to verify the experiment results in a more consistent and comparable way.

### 5.4.1 Enumeration Reflects Decision Trees in a Concise and Effective Way

When conducting experiments with the Ecoli dataset, the decision tree model resulting from the initial classification contains 7 branches and 13 nodes, as shown in Figure 5.1. On completion of its enumeration, the digitized form of branches and nodes is shown in the second row of Table 5-1.

Each node is uniquely tagged by a digital ID, and each ID reflects the node's hierarchical location in the tree. Because of its hierarchical and self-referenced nature, the ID also encapsulates its preceding nodes of the same branch and presents itself as a compact and enumerated decision path. Therefore, a collective display of each branch's leaf node resembles the decision tree model in a simplified and digitized form, as shown in the third row in this table.

Table 5-1 - Ecoli dataset's decision tree in digitized form

Branches and nodes are shown as a set of decision rules	Rule 1: $alm1 \leq 0.57$ and $aac \leq 0.64$ and $mcc \leq 0.73$ : neg Rule 2: $alm1 \leq 0.57$ and $aac \leq 0.64$ and $mcc > 0.73$ and $lip \leq 0.48$ and $aac \leq 0.51$ : neg Rule 3: $alm1 \leq 0.57$ and $aac \leq 0.64$ and $mcc > 0.73$ and $lip \leq 0.48$ and $aac > 0.51$ : pos Rule 4: $alm1 \leq 0.57$ and $aac \leq 0.64$ and $mcc > 0.73$ and $lip > 0.48$ : pos Rule 5: $alm1 \leq 0.57$ and $aac > 0.64$ and $alm1 \leq 0.3$ : neg Rule 6: $alm1 \leq 0.57$ and $aac > 0.64$ and $alm1 > 0.3$ : pos Rule 7: $alm1 > 0.57$ : pos
Enumerated map showing all node IDs in each decision path	Branch 1: 1.0 -> 1.1 -> 1.1.1 -> 1.1.1.1 Branch 2: 1.0 -> 1.1 -> 1.1.1 -> 1.1.1.2 -> 1.1.1.2.1 -> 1.1.1.2.1.1 Branch 3: 1.0 -> 1.1 -> 1.1.1 -> 1.1.1.2 -> 1.1.1.2.1 -> 1.1.1.2.1.2 Branch 4: 1.0 -> 1.1 -> 1.1.1 -> 1.1.1.2 -> 1.1.1.2.2 Branch 5: 1.0 -> 1.1 -> 1.1.2 -> 1.1.2.1 Branch 6: 1.0 -> 1.1 -> 1.1.2 -> 1.1.2.2 Branch 7: 1.0 -> 1.2
Leaf node IDs resemble a simplified and enumerated tree	Branch 1: 1.1.1.1 Branch 2: 1.1.1.2.1.1 Branch 3: 1.1.1.2.1.2 Branch 4: 1.1.1.2.2 Branch 5: 1.1.2.1 Branch 6: 1.1.2.2 Branch 7: 1.2

When conducting experiments with the Pima Indians diabetes dataset, the decision tree model resulting from the initial classification contains 20 branches and 39 nodes, as shown in Figure 5.2. After the tree enumeration process, these branches and nodes are tagged and mapped by their leaf-node IDs concisely and effectively, as shown in the second and third row of Table 5-2.

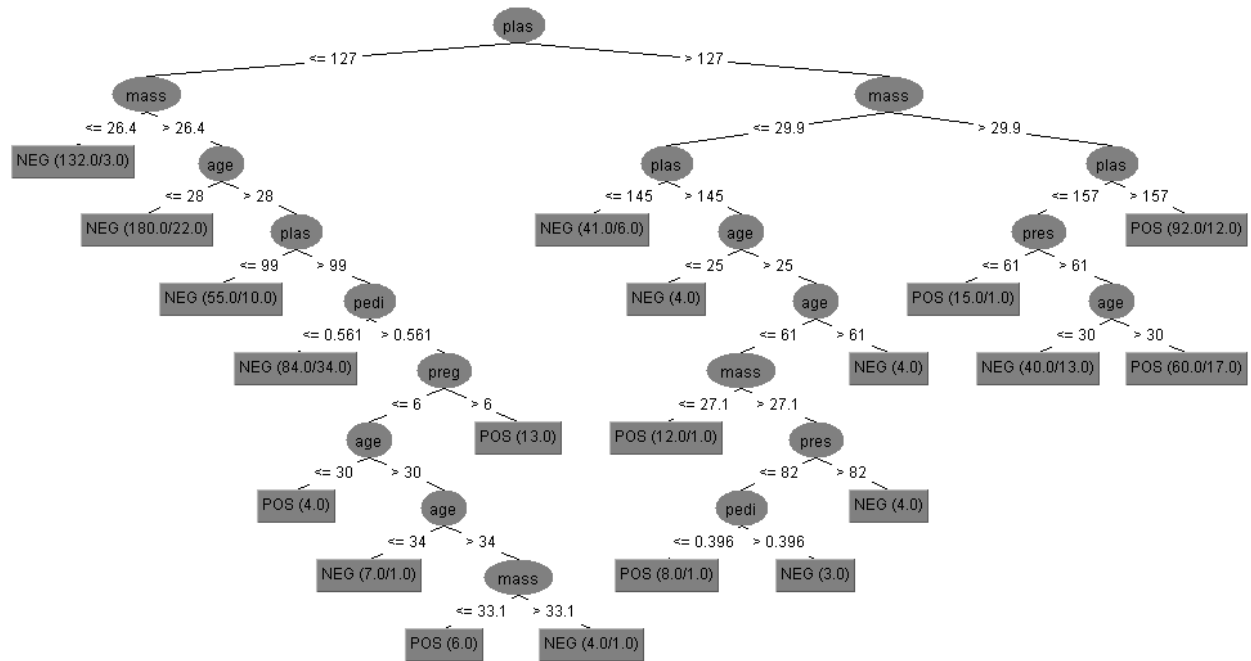


Figure 5.2 – Pima Indians diabetes dataset's decision tree model

Table 5-2 - Pima Indians dataset's decision tree represented by enumerated leaf-node IDs

Branches and nodes are shown as a set of decision rules	<p>Rule 1: plas &lt;= 127 and mass &lt;= 26.4: NEG</p> <p>Rule 2: plas &lt;= 127 and mass &gt; 26.4 and age &lt;= 28: NEG</p> <p>Rule 3: plas &lt;= 127 and mass &gt; 26.4 and age &gt; 28 and plas &lt;= 99: NEG</p> <p>Rule 4: plas &lt;= 127 and mass &gt; 26.4 and age &gt; 28 and plas &gt; 99 and pedi &lt;= 0.561: NEG</p> <p>Rule 5: plas &lt;= 127 and mass &gt; 26.4 and age &gt; 28 and plas &gt; 99 and pedi &gt; 0.561 and preg &lt;= 6 and age &lt;= 30: POS</p> <p>...</p> <p>Rule 18: plas &gt; 127 and mass &gt; 29.9 and plas &lt;= 157 and pres &gt; 61 and age &lt;= 30: NEG</p> <p>Rule 19: plas &gt; 127 and mass &gt; 29.9 and plas &lt;= 157 and pres &gt; 61 and age &gt; 30: POS</p> <p>Rule 20: plas &gt; 127 and mass &gt; 29.9 and plas &gt; 157: POS</p>
Enumerated map showing all node IDs in each decision path	<p>Rule 1: 1.0, 1.1, 1.1.1</p> <p>Rule 2: 1.0, 1.1, 1.1.2, 1.1.2.1</p> <p>Rule 3: 1.0, 1.1, 1.1.2, 1.1.2.2, 1.1.2.2.1</p> <p>Rule 4: 1.0, 1.1, 1.1.2, 1.1.2.2, 1.1.2.2.2, 1.1.2.2.2.1</p> <p>Rule 5: 1.0, 1.1, 1.1.2, 1.1.2.2, 1.1.2.2.2, 1.1.2.2.2.2, 1.1.2.2.2.2.1, 1.1.2.2.2.2.1.1</p> <p>...</p> <p>Rule 18: 1.0, 1.2, 1.2.2, 1.2.2.1, 1.2.2.1.2, 1.2.2.1.2.1</p> <p>Rule 19: 1.0, 1.2, 1.2.2, 1.2.2.1, 1.2.2.1.2, 1.2.2.1.2.2</p> <p>Rule 20: 1.0, 1.2, 1.2.2, 1.2.2.2</p>

Leaf node	Rule 1: 1.1.1
IDs resemble a	Rule 2: 1.1.2.1
simplified	Rule 3: 1.1.2.2.1
and enumerated	Rule 4: 1.1.2.2.2.1
tree	Rule 5: 1.1.2.2.2.2.1.1
	...
	Rule 18: 1.2.2.1.2.1
	Rule 19: 1.2.2.1.2.2
	Rule 20: 1.2.2.2

When conducting experiments with the Wisconsin cancer dataset, the decision tree model resulting from the initial classification contains 14 branches and 27 nodes, as shown in Figure 5.3. After the tree enumeration process, these branches and nodes are tagged and mapped by their leaf-node IDs concisely and effectively, as shown in the second and third row of Table 5-3.

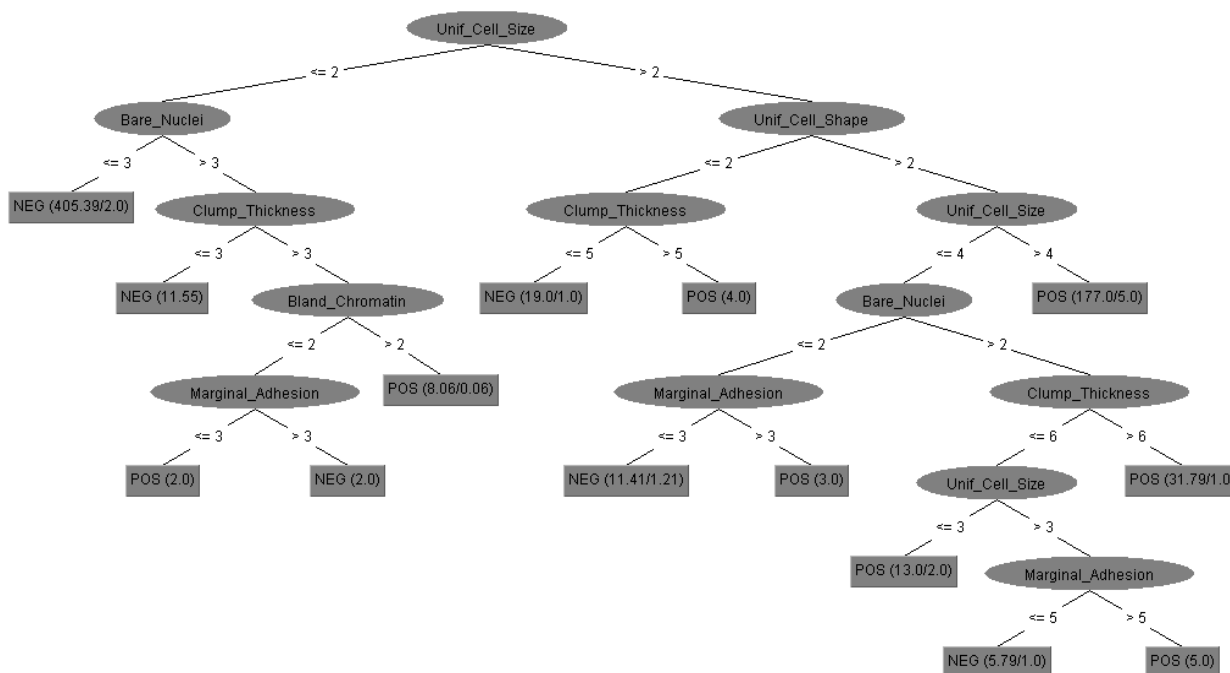


Figure 5.3 - Wisconsin cancer dataset's decision tree model



Table 5-3 - Wisconsin dataset's decision tree represented by enumerated leaf-node IDs

Branches and nodes are shown as a set of decision rules	Rule 1: Unif_Cell_Size $\leq$ 2 and Bare_Nuclei $\leq$ 3: NEG Rule 2: Unif_Cell_Size $\leq$ 2 and Bare_Nuclei $>$ 3 and Clump_Thickness $\leq$ 3: NEG Rule 3: Unif_Cell_Size $\leq$ 2 and Bare_Nuclei $>$ 3 and Clump_Thickness $>$ 3 and Bland_Chromatin $\leq$ 2 and Marginal_Adhesion $\leq$ 3: POS ... Rule 12: Unif_Cell_Size $>$ 2 and Unif_Cell_Shape $>$ 2 and Unif_Cell_Size $\leq$ 4 and Bare_Nuclei $>$ 2 and Clump_Thickness $\leq$ 6 and Unif_Cell_Size $>$ 3 and Marginal_Adhesion $>$ 5: POS Rule 13: Unif_Cell_Size $>$ 2 and Unif_Cell_Shape $>$ 2 and Unif_Cell_Size $\leq$ 4 and Bare_Nuclei $>$ 2 and Clump_Thickness $>$ 6: POS Rule 14: Unif_Cell_Size $>$ 2 and Unif_Cell_Shape $>$ 2 and Unif_Cell_Size $>$ 4: POS
Enumerated map showing all node IDs in each decision path	Rule 1: 1.0, 1.1, 1.1.1 Rule 2: 1.0, 1.1, 1.1.2, 1.1.2.1 Rule 3: 1.0, 1.1, 1.1.2, 1.1.2.2, 1.1.2.2.1, 1.1.2.2.1.1 ... Rule 12: 1.0, 1.2, 1.2.2, 1.2.2.1, 1.2.2.1.2, 1.2.2.1.2.1, 1.2.2.1.2.1.2, 1.2.2.1.2.1.2.2 Rule 13: 1.0, 1.2, 1.2.2, 1.2.2.1, 1.2.2.1.2, 1.2.2.1.2.2 Rule 14: 1.0, 1.2, 1.2.2, 1.2.2.2
Leaf node IDs resemble a simplified and enumerated tree	Rule 1: 1.1.1 Rule 2: 1.1.2.1 Rule 3: 1.1.2.2.1.1 ... Rule 12: 1.2.2.1.2.1.2.2 Rule 13: 1.2.2.1.2.2 Rule 14: 1.2.2.2

When conducting experiments with the Liver Disorders dataset, the decision tree model resulting from the initial classification contains has 26 branches and 51 nodes, as shown in Figure 5.4. After the tree enumeration process, these branches and nodes are tagged and mapped by their leaf-node IDs concisely and effectively, as shown in the second and third row of Table 5-4.



simplified and enumerated tree	Rule 24: 1.2.2.1.2.1 Rule 25: 1.2.2.1.2.2 Rule 26: 1.2.2.2
--------------------------------------	--

When conducting experiments with the Page Blocks dataset, the decision tree model resulting from the initial classification contains 41 branches and 81 nodes, as shown in Figure 5.5. After the tree enumeration process, these branches and nodes are tagged and mapped by their leaf-node IDs concisely and effectively, as shown in the second and third row of Table 5-5.

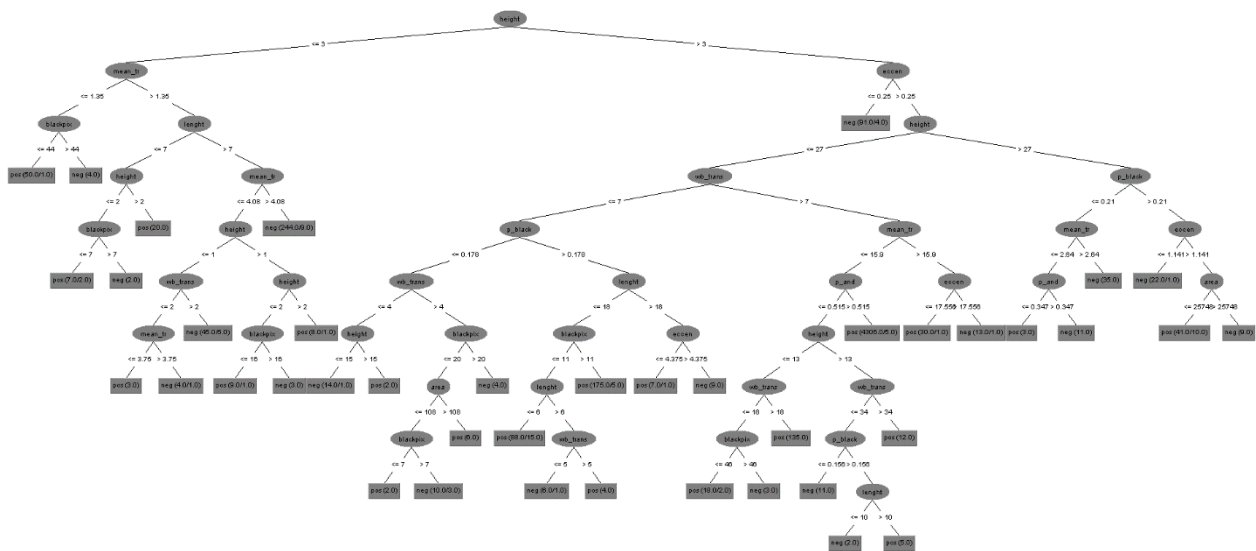


Figure 5.5 - Page Blocks dataset's decision tree model

Table 5-5 - Page Blocks dataset's decision tree represented by enumerated leaf-node IDs

Branches and nodes are shown as a set of decision rules	<p>Rule 1: height <math>\leq 3</math> and mean_tr <math>\leq 1.35</math> and blackpix <math>\leq 44</math>: pos</p> <p>Rule 2: height <math>\leq 3</math> and mean_tr <math>\leq 1.35</math> and blackpix <math>&gt; 44</math>: neg</p> <p>Rule 3: height <math>\leq 3</math> and mean_tr <math>&gt; 1.35</math> and lenght <math>\leq 7</math> and height <math>\leq 2</math> and blackpix <math>\leq 7</math>: pos</p> <p>...</p> <p>Rule 39: height <math>&gt; 3</math> and eccen <math>&gt; 0.25</math> and height <math>&gt; 27</math> and p_black <math>&gt; 0.21</math> and eccen <math>\leq 1.141</math>: neg</p>
---	---

	Rule 40: height > 3 and eccen > 0.25 and height > 27 and p_black > 0.21 and eccen > 1.141 and area <= 25748: pos Rule 41: height > 3 and eccen > 0.25 and height > 27 and p_black > 0.21 and eccen > 1.141 and area > 25748: neg
Enumerated map showing all node IDs in each decision path	Rule 1: 1.0, 1.1, 1.1.1, 1.1.1.1 Rule 2: 1.0, 1.1, 1.1.1, 1.1.1.2 Rule 3: 1.0, 1.1, 1.1.2, 1.1.2.1, 1.1.2.1.1, 1.1.2.1.1.1 ... Rule 39: 1.0, 1.2, 1.2.2, 1.2.2.2, 1.2.2.2.2, 1.2.2.2.2.1 Rule 40: 1.0, 1.2, 1.2.2, 1.2.2.2, 1.2.2.2.2, 1.2.2.2.2.2, 1.2.2.2.2.2.1 Rule 41: 1.0, 1.2, 1.2.2, 1.2.2.2, 1.2.2.2.2, 1.2.2.2.2.2, 1.2.2.2.2.2.2
Leaf node IDs resemble a simplified and enumerated tree	Rule 1: 1.1.1.1 Rule 2: 1.1.1.2 Rule 3: 1.1.2.1.1.1 ... Rule 39: 1.2.2.2.2.1 Rule 40: 1.2.2.2.2.2.1 Rule 41: 1.2.2.2.2.2.2

Once each individual node is uniquely tagged and mapped, its classification statistics can then be collected, stored and analyzed at a node- and micro-level individually, as compared to the typical model- and macro-level analysis based on the whole decision tree and its overall result is examined holistically. Some of the statistics collection and comparison results are discussed in the following section.

#### 5.4.2 Node-Level Statistics Comparison and Error-Sensitive Pattern Identification

There are different ways to examine node-level statistics, for example, by comparing the heaviest and lightest nodes using the highest and lowest counts of instances that are traversed through, however the focus of this study is to identify and explore the weakest nodes with the highest error rates and the value patterns involved.

As a first step, the attributes and values involved with the top-three nodes in the error rate ranking are compared with the top-three ranked attributes identified in the error-

sensitive attribute (ESA) evaluation to cross-check these two different error-sensitive pattern evaluation methods, as shown in Table 5-6.

It is shown that for three of the datasets, Pima Indians diabetes, Wisconsin and Page Blocks, their top ranked nodes contain all the top-ranked error-sensitive attributes, and for the other two datasets, Ecoli and Liver Disorders, their top ranked nodes contain part of the top-ranked error-sensitive attributes identified but not all.

Table 5-6 - The weakest nodes' attributes & values vs. the most error-sensitive attributes

Rank	Ecoli dataset's enumerated tree node & error-rate ... attributes & values involved	Attributes identified in ESA evaluation by attribute-error count
1	1.2 (3.36%: 4/119) ... <u>alm1</u> >0.57	chg (17)
2	1.1.1.1 (3.17%: 6/189) ... <u>alm1</u> <=0.57 & aac<=0.64 & mcg<=0.73	<u>alm1</u> (15)
3	1.1.1 (3.02%: 6/199) ... <u>alm1</u> <=0.57 & aac<=0.64	lip (14)
Rank	Pima diabetes dataset's enumerated tree node & error-rate ... attributes & values involved	Attributes identified in ESA evaluation by attribute-error count
1	1.1.2.2.2.1 (40.48%: 34/84) ... <u>plas</u> <=127 & <u>mass</u> >26.4 & <u>age</u> >28 & <u>plas</u> >99 & pedi<=0.561	<u>plas</u> (83)
2	1.2.2.1.2.1 (32.50%: 13/40) ... <u>plas</u> >127 & <u>mass</u> >29.9 & <u>plas</u> <=157 & pres>61 & <u>age</u> <=30	<u>mass</u> (70)
3	1.1.2.2.2 (30.51%: 36/118) ... <u>plas</u> <=127 & <u>mass</u> >26.4 & <u>age</u> >28 & <u>plas</u> >99	<u>age</u> (31)
Rank	Wisconsin cancer dataset's enumerated tree node & error-rate ... attributes & values involved	Attributes identified in ESA evaluation by attribute-error count
1	1.2.2.1.2.1.2.1 (20.00%: 1/5) ... <u>UC_Sz</u> >2 & <u>UC_Sh</u> >2 & <u>UC_Sz</u> <=4 & Bare_Nuc>2 & <u>Clump_Th</u> <=6 & <u>UC_Sz</u> >3 & Mg_Adh<=5	<u>UC_Sz</u> - Unif Cell Size (32)
2	1.2.2.1.2.1.1 (15.38%: 2/13) ... <u>UC_Sz</u> >2 & <u>UC_Sh</u> >2 & <u>UC_Sz</u> <=4 & Bare_Nuc>2 & <u>Clump_Th</u> <=6 & <u>UC_Sz</u> <=3	<u>UC_Sh</u> - Unif Cell Shape (32)
3	1.2.2.1.2.1 (13.04%: 3/23) ... <u>UC_Sz</u> >2 & <u>UC_Sh</u> >2 & <u>UC_Sz</u> <=4 & Bare_Nuc>2 & <u>Clump_Th</u> <=6	<u>Clump_Th</u> - Clump Thickness (27)

Rank	Liver Disorders dataset's enumerated tree node & error-rate ... attributes & values involved	Attributes identified in ESA evaluation by attribute-error count
1	1.2.2.1.2.2 (40.91%: 18/44) ... gammagt>20 & drinks>5 & drinks<=12 & sgpt>21 & <u>sgot</u> >22	<u>sgot</u> (22)
2	1.1.2.2.1.1 (38.10%: 8/21) ... gammagt<=20 & sgpt>19 & <u>sgot</u> >20 & drinks<=5 & sgpt<= 26	mcv (16)
3	1.2.2.1.2 (34.55%: 19/55) ... gammagt>20 & drinks>5 & drinks<=12 & sgpt >21	alkphos (10)
Rank	Page Blocks dataset's enumerated tree node & error-rate ... attributes & values involved	Attributes identified in ESA evaluation by attribute-error count
1	1.2.2.1.1.1.2.1.1.2 (30.00%: 3/10) ... height>3 & <u>eccen</u> >0.25 & height <=27 & wb_trans<=7 & <u>p_black</u> <=0.178 & wb_trans>4 & blackpix <=20 & area<=108 & blackpix>7	<u>mean_tr</u> (89)
2	1.1.2.1.1.1 (28.57%: 2/7) ... height<=3 & <u>mean_tr</u> >1.35 & lenth<=7 & height<=2 & blackpix<=7	<u>p_black</u> (43)
3	1.1.2.2.1.1.1.2 (25.00%: 1/4) ... height<=3 & <u>mean_tr</u> >1.35 & lenth> 7 & <u>mean_tr</u> <=4.08 & height<=1 & wb_trans<=2 & <u>mean_tr</u> >3.75	<u>eccen</u> (29)
3	1.2.2.1.1.1.2.1.1 (25.00%: 3/12) ... height>3 & <u>eccen</u> >0.25 & height<= 27 & wb_trans<=7 & <u>p_black</u> <=0.178 & wb_trans>4 & blackpix<= 20 & area <=108	

In the second step, data samples associated with the weakest nodes identified by the enumeration method are filtered out and a re-classification is performed. In addition, another re-classification for the data samples with the most error-sensitive attributes filtered out is also performed, so the two sets of re-classification results can be compared and evaluated.

The initial results confirm improved classification accuracy in all five datasets after their weakest records are filtered out based on their respective weakest node in the tree, and one dataset improves significantly, as shown in Table 5-7.

This table also shows the scenario re-test results when filtering out the most error-sensitive attribute, where three datasets show better accuracy and the other two datasets show worse results. Amongst the three-way test results, the cells with a green ticked-box (✓) are the ones with improved accuracy when compared to the baseline results in the first column, and the cells with a red cross (✗) are the ones suffering lower accuracy, and the error-sensitive attributes (ESA) involved in these tests are also highlighted in red. Further analysis on the results is discussed in Section 5.5.

Table 5-7 - Three-way comparison - original vs weakest node vs most error-sensitive attribute

Ecoli's initial result with original 336 records	Re-test 217 records after removing 119 "weakest" records	Re-test original data after removing top most ESA – <b>alm1</b>
Accuracy: 94.64% with 18 errors	Accuracy: 97.70% with 5 errors ✓	Accuracy: 92.86% with 24 errors ✗
Pima diabetes' initial result with original 768 records	Re-test 684 records after removing 84 "weakest" records	Re-test original data after removing top 2 most ESAs – <b>plas</b> & <b>mass</b>
Accuracy: 73.96% with 200 errors	Accuracy: 77.19% with 156 errors ✓	Accuracy: 67.84% with 247 errors ✗
Wisconsin cancer's initial result with original 699 records	Re-test 694 records after removing 5 "weakest" records	Re-test original data after removing top 2 most ESAs - <b>UC_Sz</b> & <b>UC_Sh</b>
Accuracy: 94.13% with 41 errors	Accuracy: 95.97% with 28 error ✓	Accuracy: 95.71% with 30 errors ✓
Liver Disorders' initial result with original 345 records	Re-test 301 records after removing 44 "weakest" records	Re-test original data after removing top 2 most ESAs – <b>sgot</b> & <b>mcv</b>
Accuracy: 67.54% with 112 errors	Accuracy: 77.08% with 69 errors ✓	Accuracy: 71.01% with 100 errors ✓
Page Blocks' initial result with original 5,473 records	Re-test 5463 records after removing 10 "weakest" records	Re-test original data after removing top most ESA – <b>mean_tr</b>
Accuracy: 97.22% with 152 errors	Accuracy: 97.36% with 144 errors ✓	Accuracy: 97.24% with 151 errors ✓

## 5.5 Experiment Analysis

The purpose of decision tree enumeration is not simply to convert a graphical decision tree into a digitized map of nodes, but rather to use such a digitized map to facilitate the collection of node-level statistics for the purpose of node-level error-sensitive value pattern analysis to highlight the potentially weakest part of the decision tree which may be linked to some specific error-sensitive attributes and values to distinguish and identify data records with such risky value patterns for further error analysis and the development of error-reduction measures.

The results from the initial experiments appear to support this decision tree enumeration idea and the identification of the weakest node. While the performance improvement from re-testing the datasets without their weakest samples is consistent in all five datasets, various ways to utilize the weakest value patterns for optimal error measure and further knowledge extension are still being developed. The following sections discuss the results and possible implications.

### 5.5.1 Digitized Node IDs Facilitate Node-Level Analysis

This decision tree enumeration method makes node-level analysis easy by formulating individual node IDs in a unique, numerical and contextual way. For example, the node ID 1.1.2.2.2.1.1 in the decision tree for the Pima Indians diabetes dataset as shown in Table 5-2 is in a numeric text string format and is incorporated with its preceding node IDs hierarchically within the same branch starting from the root. Because each ID is unique to the node in the tree, classification statistics can subsequently be collected and stored for individual nodes using their IDs as the keys and later to locate and retrieve the node-level



statistics more efficiently than using the branch and node description text, e.g. “plas  $\leq$  127 and mass  $>$  26.4 and age  $>$  28 and plas  $>$  99 and pedi  $>$  0.561 and preg  $\leq$  6 and age  $\leq$  30”, even if such lengthy verbiage is consolidated and simplified as “(plas  $>$  99 and  $\leq$  127) and (age  $>$  28 and  $\leq$  30) and pedi  $>$  0.561 and preg  $\leq$  6 and mass  $>$  26.4”, however, using such text as key for information retrieval can still be awkward. As decision trees grow bigger, such concise node IDs can become more useful because of their systematic and self-referential characteristics.

One way to utilize such node-level statistics is to rank their classification error rates from high to low and the top node with the highest error rate may then be considered as the weakest node in the tree and the attributes and values associated with the weakest node may be considered more error-sensitive than the others, in relative terms.

### **5.5.2 Examining the Weakest Nodes and Error-Sensitive Value Patterns**

To evaluate the effectiveness of this decision tree enumeration method, the associated node-level classification statistics can be validated by a performance comparison, and one practical but rather non-deterministic measure is to compare the classification accuracy after some simplistic error-reduction measures are applied. For example, by using the attribute and value patterns associated with the weakest node as selection criteria, the potentially weakest and most error-sensitive data samples, including both the misclassified and correctly classified data instances, can be identified and separated for further examination and the original dataset becomes smaller in size but potentially higher in reliability and accuracy. The experiment results seem to confirm the validity of the weakest node and the

associated error-sensitive value patterns in all five datasets, some with significant improvement in accuracy and others with a modest but consistent level of improvement.

One test case which has a significant improvement is the Wisconsin cancer dataset with 699 records. After sorting and ranking the classification error rate of individual nodes, node 1.2.2.1.2.1.2.1 (20.00%: 1/5) is identified as the weakest node, as shown in Table 5-6. When the five data samples associated with this weakest node are identified and separated from the dataset, that is  $5/699=0.007\%$  reduction in sample size, accuracy improves from the initial result of 94.13% with 41 errors compared to the re-test result of 95.97% with 28 errors which is an almost 2% improvement in the re-test, and instead of one less error from the filtered out five samples, the total error count reduces from 41 to 28, as shown in Table 5-7. In the re-test for filtering out the two most error-sensitive attributes, accuracy also improves to 95.71% with 30 errors.

The other four datasets also show various levels of success in accuracy enhancement. For example, in the Liver Disorders dataset with 345 records, node 1.2.2.1.2.2 (40.91%: 18/44) is identified as the weakest node, as shown in Table 5-6, and there are 44 samples associated with this node and 18 of them are errors. A re-test to the updated dataset after filtering out these 44 weakest samples obtains an improvement in accuracy from the initial result of 67.54% with 112 errors to the re-test result of 77.08% with 69 errors, so the error count reduces by 43 instead of 18 from the filtered out samples.

One possible explanation for this noticeable result is the inclusion of the weakest node and its data samples in the initial classification process could have had a potentially adverse impact on the information gain (entropy) calculation when constructing C4.5/J48

decision trees because of their error-sensitive attributes and values, which may lead to error-prone split point conditions and the subsequent weakest nodes in the tree. If the impact of the weakest node and its data samples is significant, a change in the re-test result after filtering out the weakest samples would also likely be noticeable. If the impact is less significant, then the difference between the original and re-test result may not be so noticeable, as shown by the Page Blocks dataset in Table 5-7.

This reasoning may also partially explain why ensemble tree models, such as random forest, are considered superior to standalone tree models. The random forest model selects a portion of the data attributes randomly and generates hundreds of thousands of trees accordingly, and then votes for the best performing one to produce the classification result. The random attribute selection process may inadvertently generate and vote for trees without some highly error-sensitive attributes, and also with bigger value ranges to split due to the fewer attributes involved, therefore enabling ensemble models to produce more accurate results, but at the expense of consuming extra resources and making it more difficult to track.

### **5.5.3 Possible Contributions, Effectiveness and Weakness**

This decision tree enumeration and node-level examination idea can be considered as another step forward in the evaluation study on error-sensitive attributes discussed in Chapter 3. It expands from attribute-level evaluation to node-level and split-point value range evaluation. While still at an early stage, the experiments on this expansion produce some encouraging and consistent results, showing how the weakest tree node and associated data samples can be identified in an efficient and systematic way, proving how some of these

weakest records play a significant role in error reduction and to help make stakeholders more aware of the weakest and error-sensitive value patterns as a way to improve understanding about the correlation between errors and value patterns.

In terms of effectiveness, this enumeration process generates a unique and concise digital ID tag for each individual node of a decision tree with a contextual reference, simplifies node-level statistics collection and analysis, and expands the typical tree-level “macro” analysis with focus on the whole classification model into the node-level “micro” analysis with focus on specific attributes and value ranges in a systematic and transparent way.

It is hoped that this decision tree enumeration process can ignite more interest in transforming other classification models from a black-box style classifier into a more transparent and more traceable white-box model to help understand how each data sample can be assessed at each key decision points and potentially lead to more awareness of the specific correlation between attribute value patterns and errors and how the associated classification rules may be improved for better performance with such further information about the attribute value patterns and detailed rule formation.

While the intention of this decision tree enumeration process is reasonable and its development has been tested on some real-world datasets with limited success, questions and doubts about this approach are also obvious. First, the successful experiments are based on the removal of data samples associated with the weakest tree branch having the highest error rate, which may seem drastic and lacking in a formal and theoretical proof; however,

this highlights the usefulness and potential significance in identifying samples with the weakest value patterns.

This subsequently led to the second major issue, that is, it is unclear what should be done with the data samples associated with the weakest tree branch with the highest error rate in the next step. Their removal from the dataset improves the overall accuracy, so a new question is, should a separate model be used to evaluate these filtered-out error-sensitive samples? If the “one size fits all” approach is not recommended for the original dataset, then why not introduce a separate model for the weakest and error-sensitive samples to suit their own specific characteristics?

The third major issue is, this study is not based on ensemble methods, yet ensemble trees are now the preferred classification model due to their superior performance to standalone decision tree models. This makes the proposed decision tree digitization method less relevant to the latest classification development. This leads to the question, can some of the current developments be utilized to enumerate the transition and classification process for ensemble trees? It may not be easy to digitize and visualize the randomization and selection process of various samples and attribute combination-based parameter configuration and the circulating flow of repetition, comparison and progression. However, it is possible albeit it difficult, so this could be a direction to consider in the next phase of development.

There are still more issues to discuss which will be used as a form of inspiration to broaden and advance this study.

## 5.6 Related Work

The enumeration and digitization of decision trees for node-level statistics analysis and weakest node identification can be regarded as the integration of two areas of development: (1) the digitization of the classification model for process-path mapping; and (2) the evaluation of error-sensitive patterns and data quality for better data understanding. There has been a plethora of research work on these two separate areas, but no significant effort has been observed in uniting these two areas into one consolidated development for a better understanding of data and their specific error-sensitive value patterns in the context of a classification path and structure.

### 5.6.1 Study of Classification Model Digitization

One research focus of digitizing classification has been on formulating a system representation theory and adding materialization in a tangible and practical form to the otherwise conceptual and intangible classification models and information systems to enable such models and systems to represent the real world and data with quality and by taking on data classification models as pilot studies [Leo10] [SNG13] [PW08]. For example, some researchers define and measure some perceptible and traceable elements with some physical and distinguished meaning within classification systems rather than conceptual terminologies and abstract theories to establish a more meaningful connection between the physical world of materials and the conceptual platform of theories to cement the gap between the traditional environment and the technological development assisted by expressive and helpful guardrails and guidelines of new definitions.

*Practical instantiation* and *significance* are two such new and materialized concepts that have been studied and proposed [Leo10]. The practical instantiation of an algorithm or software can be considered as material because the abstract and conceptual ideas of an algorithm or software can turn into physical and tangible results by feeding through experiments or implementation by some specific instances such as data samples and other components. The significance of an algorithm or software process can also be considered as material because the potential consequences from abstract and conceptual ideas may still have a physical and influential impact on the environment and society, people and organizations.

In order to improve classification accuracy and data pattern comprehensibility, some researchers suggest the use of negotiations by human actors as a part of the materialization process when digitizing and representing classification systems [SNG13], and such negotiations are associated with activities such as instantiation, re-narration and meta-narration, to digitize and outline the processing paths of classification systems which are typically packaged inside a black-box in a transparent and interactive way, and also with specific materials to reflect upon and also to apply by using practical samples and scenarios. While instantiation refers to the requirements of real-world data samples and the flexibility in classification rules, both re-narration and meta-narration refer to the need for ongoing negotiation, revision and manipulation of classification process to accommodate the dynamic and sporadic nature of real-world data.

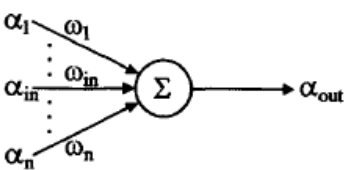
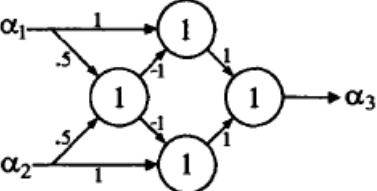
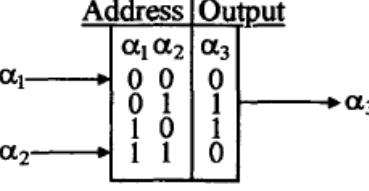
In term of decision tree classification systems, the above emphasis on interactive negotiation and modification while digitizing classification systems may partially address

the well-recognized over-fitting and under-fitting issues with decision tree models, and may also reflect the one key objective of the study reported in this chapter, that is, to enable the collection of classification statistics at a detailed node-level and to provide specific weakest nodes and error-sensitive pattern information as a kind of substantial material to facilitate some meaningful negotiation and enhancement modification.

### 5.6.2 A More Practical Digitization Example

Another practical example is the digitization of artificial neural networks [Bad94]. For artificial neural networks that are created under the McCulloch-Pitts neuron model, when the inputs of a neuron with their respective weights are aggregated and regulated by a threshold value function, they can then be considered as a binary type of on (1) and off (0) value, a transformation from analog signals into digital form, as shown in the simplified digitization process in Table 5-8.

Table 5-8 – Basic process in digitizing artificial neural networks

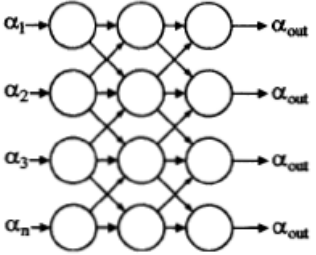
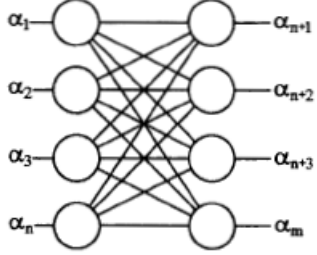
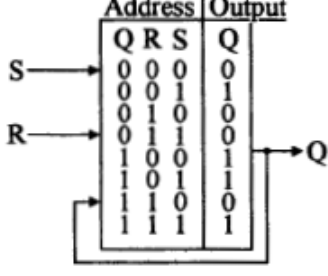
		
<p>McCulloch-Pitts neuron model:  <math>\alpha</math> = activation, <math>\omega</math> = weights, <math>\Sigma</math> = activation threshold function</p>	<p>Apply XOR function on weights and threshold in the neuron</p>	<p>A simplified digital neuron with values of the truth table for XOR function</p>

When more layers of neurons are added to the network, digital feedback can also be simulated by including latch functions to contribute to the learning and back-



propagation process and to enable neurons from the feed-forward mode to bi-directional mode, as shown in Table 5-9.

Table 5-9 – Digital feedback by latches as a form of back-propagation

		
Mapping input and output of a feed-forward neuron network	Bi-directional neuron network model needs feedback	Feedback in digitized neurons is controlled by R-S latch - S = set, R = reset, Q = output

While this artificial neural network digitization process has been used as a reference point in an electromechanical implementation for signal errors and short-circuit diagnosis for power generators [TCC13], it has not been mentioned or tested in other data mining and classification research work so far.

It is admirable and inspiring to see the genuine effort and robust discussions on the need for and benefits of how best to define and divide the abstract and conceptual algorithms and systems from the real-world data and environment, and then to add more definitions and integrated research on how to connect the two different worlds, separated by earlier studies, one categorized as abstract and one categorized as material, back together.

When working on data mining and classification tasks, sometimes it is necessary to divide and discuss conceptual algorithms and real-world data separately and theoretically to pursue scientific clarity and validity. At other times, it is important to examine and

evaluate the algorithms and data in a systematic and correlated way to obtain a practical solution and result. Hence, it is situational.

The idea of decision tree digitization with each node uniquely tagged and tracked with classification statistics can be considered as an attempt to materialize the conceptual classification model and the identification of the weakest node and error-sensitive value patterns can be considered as the practical instantiation of real-world data. Therefore, this independent study may have coincidentally developed and applied what Leonardi and other researchers [Leo10] [SNG13] [Bad94] have studied and theorized in recent years, but in a very limited way and with its current focus on the decision tree classification model. The intention of this study is to expand this digitization trial to the random forest classification model in the next stage of exploration.

## 5.7 Summary

The work reported in this chapter addresses the question - “Is there a way to identify the Achilles’ heel of a classification model?”, that is, finding a way to locate the weakest and most error-sensitive part of a model. To achieve this goal, this study developed a decision tree digitization method to facilitate the identification and examination of the potentially weakest nodes and error-sensitive value patterns in the model using decision trees as a pilot model. Initial experiments demonstrated more meaningful and successful results when compared to earlier evaluation studies of error-sensitive attributes.

It has been recognized that a digitization process for tagging and mapping key critical decision points along each classification path can be as important as marking the weakest points in the model. Such a digitized mapping process can potentially help users and

stakeholders better understand the specific flow of steps and the conditions of value patterns involved in a classification process for a particular dataset. In addition to transforming a classification model which normally works like a black-box into a map of workflows with classification statistics collected at each key decision point, it can also help review how specific value patterns are correlated with classification results and which specific path or part of the classification model may be more effective or more error-sensitive. It is more about tagging and showing the workflow of the classification model with details as a form of knowledge development about the model in the context of value patterns and the data as a whole.

This development has subsequently led to a more context-focused study by advancing to a new stage of the research journey as discussed in the next chapter.

## CHAPTER 6

### Transformation of Confusion Matrices for Contextual Data Analysis

#### 6.1 Introduction

Context consideration can be important to help explore understanding and decision making. The inclusion of less documented historical and environmental contexts in researching diabetes amongst Pima Indians by Schulz et al. and Phihl [SBR06] [Pil12] uncovered reasons which were more likely to explain why some Pima Indians ‘have up to ten times the rate of diabetes as Caucasians’ (Phihl, 2012. p.2), finding that the reasons were not due to their specific genetic patterns or ethnicity, but “because of environmental, social, economic and historical causes.” (Phihl, 2012. p.2) Phihl’s study also suggested that “the majority of the results showed that genetics are a contributing factor ... (because) genetic studies have been well funded through America’s capitalist driven healthcare systems.” (Phihl, 2012. p.12) This claim of new understanding is thought-provoking but requires more evidence and stronger proof.

If historical and environmental factors are considered as external contexts when not included as part of a dataset for research, some forms of internal contexts may also exist inside the dataset without being declared. This chapter discusses a context construction model that transforms a confusion matrix from a classification result table into a matrix of categorical, incremental and correlational contexts to emulate a kind of internal context as a way to explore and expand understanding about data and results.

Initial experiments on binary classification scenarios revealed some interesting findings, while some of the experiments and results have been summarized in one earlier

reports focusing on weakly supervised learning [Wu19a], more details about this exploration process are further analyzed in the next few sections. When negative and positive instances are compared to happy families and unhappy families, this resulted in contexts reflecting the Anna Karenina principle well, that is - happy families are all alike; every unhappy family is unhappy in its own way, which is an encouraging sign for further contextual analysis into the details of the hows when planning for a happy family or a data classification task in a world of uncertainties.

The rest of this chapter is organized as follows. Section 6.2 provides the reasons to initiate this context construction and matrix transformation idea and outlines its potential benefits. Section 6.3 outlines the key steps involved in this contextual data analysis. Section 6.4 summarizes the experiments and results on the selected datasets. Section 6.5 discusses potential problems and issues. Section 6.6 reviews some early influential work that inspired this study. Finally, Section 6.7 concludes the development of this contextual model and outlines a plan for future exploration.

## **6.2 The Reasons and Potential Benefits of Transforming a Confusion Matrix**

### **6.2.1 Why Context is Important in Error and Data Analysis?**

In addition to examples mentioned in Chapter 1 and Chapter 2, another report [Wu19b] has demonstrated the importance of context consideration is the detection of the fraudulent financial data patterns of Enron in data mining research based on publicly available financial data [WHM12]. Enron, an American energy company, declared bankruptcy in December 2001 and its share price dived from \$90 in August 2000 to \$0.12 in January 2002.

Its collapse was due to financial fraud committed by the management team, which resulted in the loss of thousands of jobs and the demise of Arthur Andersen, one of the largest accounting firms in the world at the time [Seg19].

It is widely accepted that the identification of management fraud can be difficult because high-level management authorities can easily overrule internal controls and protocols to falsify reports with a high level of sophistication and collusion. However, by comparing the reported data from a fraudulent firm like Enron with contextual information, such as a "centroid" model from an aggregated and balanced data set of industry-representative firms, together with finer-grained historical data integration and comparison based on quarterly reports in addition to yearly reports, some unusual patterns became more conspicuous in Enron's data within the context of "centroid" values, especially when compared progressively in a quarter-by-quarter way, as shown in Figure 6.1.

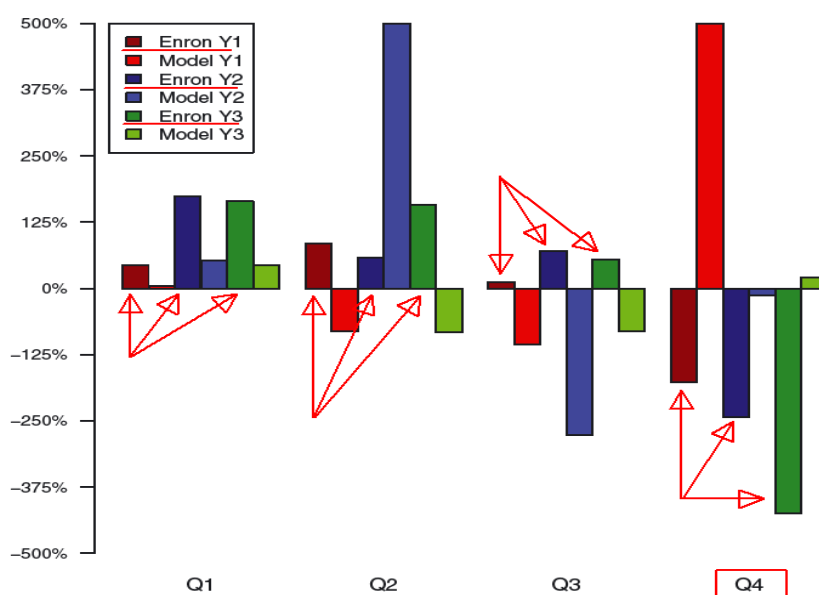


Figure 6.1 - Quarterly cash flow earnings ratio comparison between Enron and its industry model

This figure shows that Enron's cash flow earnings ratio increased as the fraud progressed in its final three years. While showing increases in the first three non-audited quarters of each year, each fourth quarter turned sharply negative when it was audited as part of the year-end activities and with more special purpose entities to be included or manipulated in the annual balance sheet.

On the other hand, unusual financial data patterns do not necessarily indicate a certain fraud, as shown in Figure 6.2. While Enron's Year 2000 revenue was showing too good to be true and it indeed committed management fraud, it would be wrong to speculate Texaco and Goldman Sachs were also likely associated with management fraud simply because they were showing similarly and exceptionally good performance but on a smaller scale.

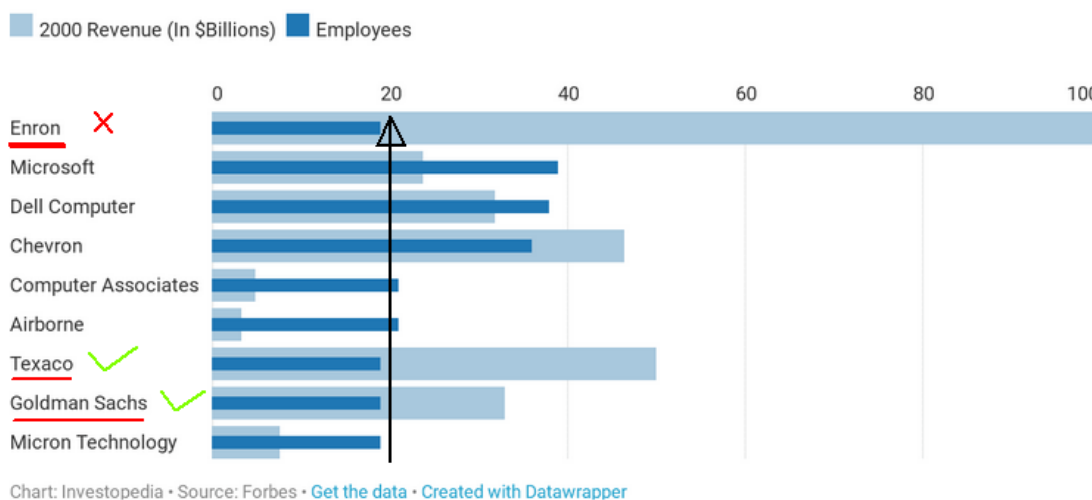


Figure 6.2 - Enron's Year 2000 Reported Revenue vs. Similarly Sized Companies

Nevertheless, wrong decision and errors are inevitable, so instead of attempting to avoid them or get rid of them at all costs, it may be more sensible to look into the errors closely in order to understand them better, to analyze errors and value patterns from various

perspectives, to examine and evaluate errors and value patterns rationally, systematically and contextually; in essence, and within context, to consider errors as a part of the knowledge in order to help develop better preventive and corrective measures.

### **6.2.2 Why Transform a Confusion Matrix into a Matrix of Context?**

On the other hand, a confusion matrix is a simple and effective way to summarize classification results and algorithm performance in a tabular form by tallying the category results. It can be easily constructed based on the categorized classification results and does not depend on any specific classification algorithm. In fact, standard data mining tools, such as WEKA and R, already include a confusion matrix as a built-in feature of their classifier packages.

A confusion matrix also makes result comparison between class label categories easy, especially for the binary classification scenarios used in this study (true negatives and true positives), and also for error categories (false negatives and false positives). It is suggested in this study that such a cross-category comparison can be viewed as a form of context for individual category statistics, in conjunction with other performance statistics such as accuracy and sensitivity.

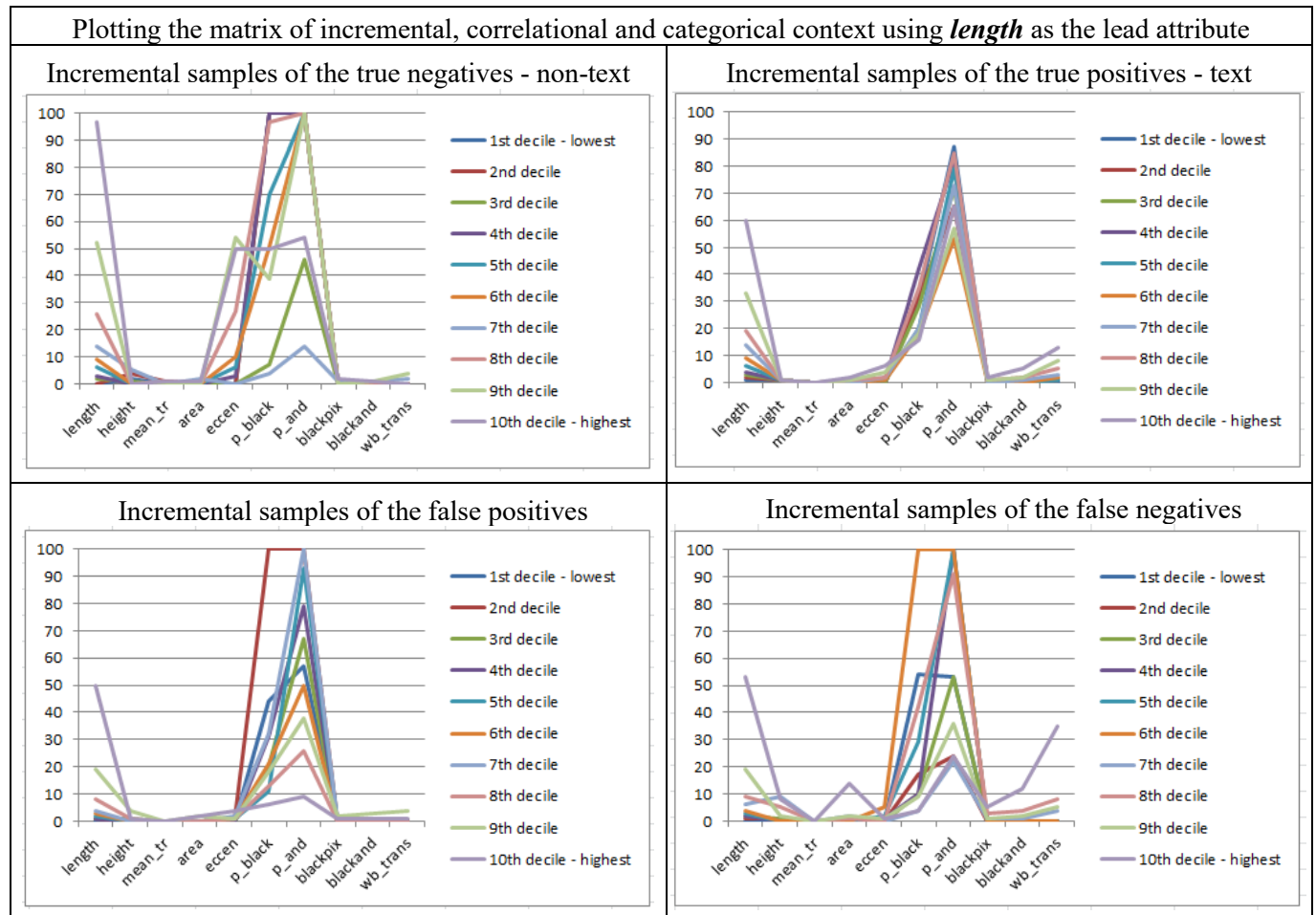
Because of these advantages, a new use is proposed for the confusion matrix, which is to transform the confusion matrix into a contextual matrix for data analysis and comparison in a multi-dimensional and comparative way, and each table cell that represents a result category also depicts the incremental and correlational context between a lead attribute and the other attributes belonging to the same result category, as shown in Table 6-1,



and more details about the construction process of this table are explained later in Section

#### 6.4.4 Constructing Contexts for Other Datasets with reference to Table 6-9.

Table 6-1 – A Confusion matrix of contexts constructed for the Page Blocks dataset



### 6.2.3 Key Components of the Confusion Matrix

The current confusion matrix transformation process is basically to convert the classification summary table into a matrix of incremental and correlational context and the resulting categories become the basis for the categorical context. The term *incremental context* describes a situation in which data samples are sorted in an ascending order to represent the

value growth path of a lead attribute selected from a rotation list of attributes. A rotation list can either be a full or partial set of attributes in their natural order in the original dataset. It can also be a full or partial set of attributes prioritized in a certain order, such as their ranked level of significance in terms of information gain or gain ratio, etc., so selected or significant attributes can be processed and compared accordingly. The term *correlational context* describes the subsequent and correlated value variation of the other attributes along the value growth path of the lead attribute.

To simplify the presentation of the value growth path of a lead attribute, ten sorted deciles of its values are used to represent ten key growth stages as incremental context in a vertical form. Starting from each key growth stage of the lead attribute, plots that are connecting across with other attributes of the same instance record can be considered as the correlational context in horizontal form. The matrix structure itself, which hosts and displays these two contexts, can then be considered as a third form of context, the categorical context.

#### **6.2.4 Potential Benefits of Confusion Matrix Transformation**

When constructing relevant data contexts, such as incremental and correlational context, and housing them in the structure of the confusion matrix, this implicitly adds another dimension of context, the categorical context, onto the analytic model and makes a cross-category comparison between contexts simple and systematic, category-by-category in a side-by-side manner, all within a simple matrix structure.

For example, the observation of diverging points and distinctive correlation patterns along the value growth path of a lead attribute between result categories may add to

the understanding of which point of the value range of this lead attribute may be more likely to be involved in the separation between true negatives and true positives, and in confusion leading to errors. Likewise, the observation of converging points and correlation patterns along the value growth path of a lead attribute between categories may help explore patterns of ambiguity between class labels and errors and may lead to further discussion on data centric questions such as:

- Does the classification rate of positive or negative instances increase when the value of an attribute increases?
- Do value patterns of type-1 (false positive - FP) and type-2 (false negative - FN) errors vary accordingly when the value of an attribute increases?
- Are there transitional value patterns or threshold value patterns between negative and positive categories, or between type-1 and type-2 errors, when the value of an attribute increases?

Questions like these may sound simplistic and trivial, but they represent some genuine interest in understanding the data and the correlation between a lead attribute's value growth path and its growth impact on other attributes and the classification result and may subsequently lead to more specific questions and investigation, and therefore a better understanding of the data.

Some standard data mining software such as WEKA has implemented features to establish a specific categorical and incremental context for individual attributes. In the example of the Pima Indians diabetes dataset, the class label distribution of the five selected attributes, *plas*, *insu*, *mass*, *pedi* and *age*, can be illustrated individually under each selected attribute in WEKA, as shown in Figure 6.3.

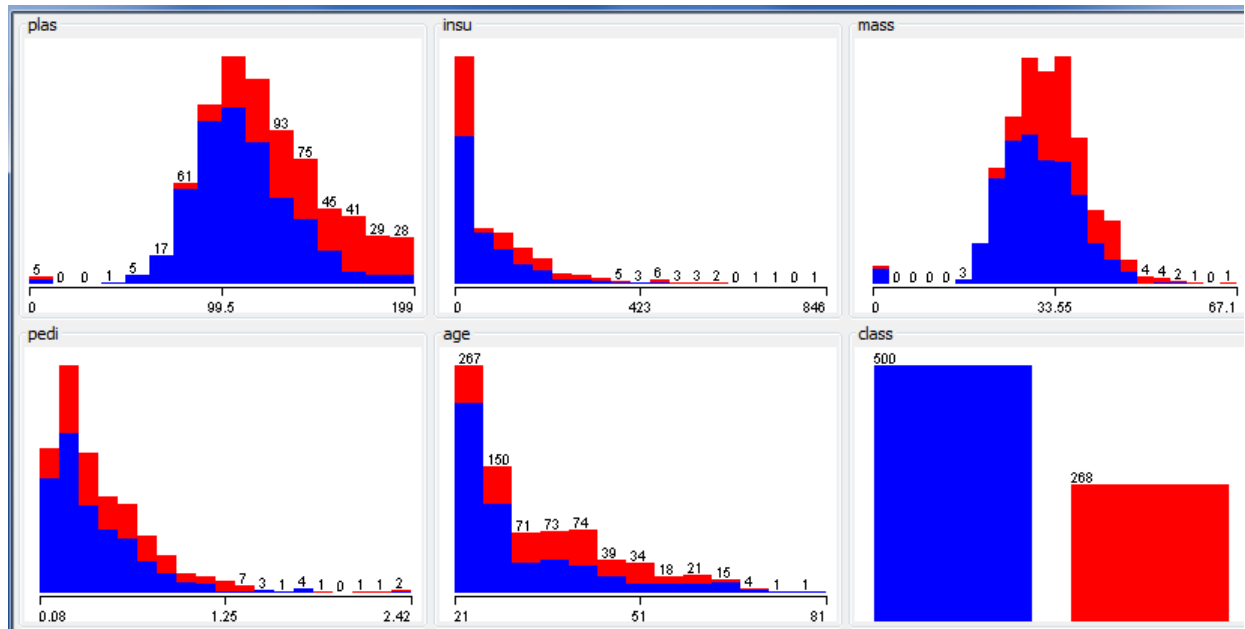


Figure 6.3 - Class distribution of individual attributes in Pima Indians diabetes dataset

One obvious issue with this individualized approach is the lack of integration to consolidate multiple key attributes into one correlated background; therefore, the association of class labels and the value growth of an attribute cannot be easily mapped and related to other attributes in a systematic way. Moreover, this illustrative feature of WEKA is based on individual attributes and is also offered as data preparation processes, consequently, similar context illustration is not easily available for misclassified error samples.

As a way to contribute to the integration of this attribute-by-attribute correlation analysis, the next section discusses the details and steps involved in establishing a multi-level context model for data analysis based on a data classification process.

### 6.3 Context Construction with Confusion Matrix

#### 6.3.1 Post-Classification Analysis and Data Normalization

This context construction can start as a part of the post-classification analysis process, and the underlying classification platform can be based on any classification algorithm, e.g. decision tree, neural network or naive Bayes, as illustrated in Figure 6.4.

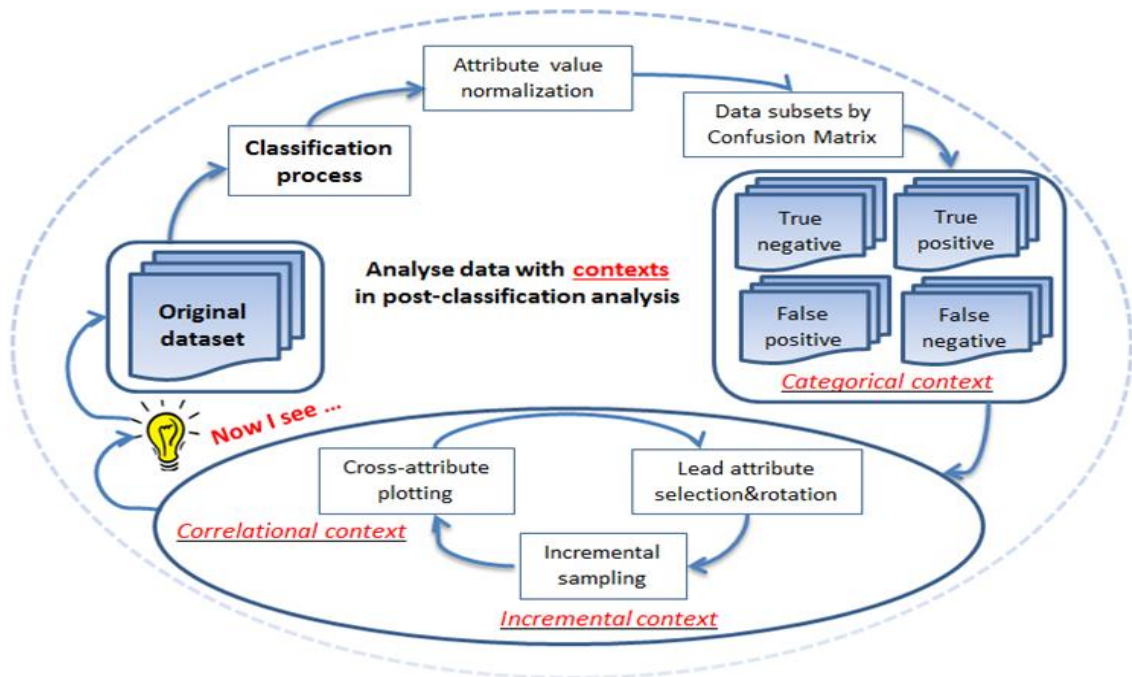


Figure 6.4 – Context construction using a confusion matrix during post-classification analysis

The resulting confusion matrix can be evaluated first and if suitable, be transformed into a table structure with each of its category cells to house specific contexts of data samples in the same category accordingly.

To construct an incremental and correlational context for an easy and consistent comparison between attributes in a dataset, values from different attributes measured in

different units are converted into one standardized value range between 0 and 100 via a min-max normalization routine:

$$NormVal = \frac{CurrVal - AttrMinVal}{AttrMaxVal - AttrMinVal} * 100 \quad (6.1)$$

This min-max normalization can make cross-attribute value plotting and comparison clearer and simpler, but it can also be problematic when dealing with skewed and outlier values. These issues are discussed in Section 6.5.

### 6.3.2 Construction of Categorical, Incremental and Correlational Context

After the involved attribute values are normalized, categorical context can be constructed by redistributing the normalized data values into separate data subsets according to their classification results as categorized in the resulting confusion matrix to allow for a visual and intuitive comparison between a specific data record and its peers within the same result category and also between other result categories, especially between the false negatives and false positives. These two error categories help identify the distinctive value patterns between result categories as a way for categorical context analysis.

Incremental context can be constructed in three steps. First, a lead attribute is selected and a common attribute selection sequence is also shared by data subsets in all result categories. Next, the values of the lead attribute in each subset are sorted sequentially and divided into ten deciles. Then, the median record from each decile is added into an incremental sample set and these ten samples from the ten deciles can be considered as a kind of reflection of ten incremental growth stages between the minimum and maximum values

of this lead attribute to help measure the incremental growth pace and gap of the lead attribute as a form of incremental context in a limited scope.

Correlational context can subsequently be constructed by plotting the incremental sample values of a lead attribute from each of its growth stage and across other attributes of the same sample record. Such cross-attribute graphs along the value growth path of the lead attribute may provide an indication on how the lead attribute may influence or be associated with other attributes in a correlational way based on its value growth trend and any significant variation in value patterns between the result categories may be highlighted as an area of interest that is warranted for a further and localized examination.

The process flow of this context construction model can be simplified and summarized into the following six-step procedure:

1. Evaluate initial classification result and its confusion matrix
2. Define a selection list of lead attributes and a rotation sequence which can be by their significance level or other criteria
3. Normalize attribute values into one standardized value range between 0 and 100 via the min-max normalization routine
4. Sort values of current lead attribute from low to high and redistribute data records into categorized subsets according to their confusion matrix categories
5. For each categorized subset sorted by the current lead attribute
  - divide subset records into ten deciles and extract median record from each decile to simulate ten key value growth stages in a vertical way within its categorized cell as incremental context
  - start from the current lead attribute and for its ten growth stages, connect and plot lines across other attributes from the same sample instance in a horizontal way as correlational context
  - compare these incremental and correlational contexts in the form of value and line patterns between categories as categorical context
6. Select the next lead attribute from the rotation list determined in Step-2 and repeat Steps 4 to 6 to construct another contextual model for the new lead attribute with its own incremental, correlational and categorical context

If significant patterns can be observed between matrix categories for a lead attribute, it can be an indication that some specific characteristics of this lead attribute and value patterns may influence or be associated with the underlying result categories and may be worth further exploration as to why and how this disparity occurs between result categories within the specific multi-dimensional contextual environment. Such contextual analysis may help raise the attention of stakeholders and domain experts with specific value patterns and contextual information and may potentially lead to better understanding and further knowledge discovery about the data.

## 6.4 Experiments and Analysis

Experiments with ten UCI datasets produced some encouraging results which support this contextual analysis idea and part of these experiments and results are summarized in the following sections.

### 6.4.1 Utilization of Existing Classification Models and Results

The initial classification result and confusion matrix for the experiment with the Pima Indians diabetes dataset of eight attributes and 768 records are shown in Table 6-2.

Table 6-2 - Classification result for Pima diabetes dataset and options for lead attributes

Correctly Classified 567 at 73.83 % Incorrectly Classified 201 at 26.17 %	Top-3 attributes ranked by gain ratio evaluator	Top-3 attributes ranked by attribute-error counter
== Confusion Matrix == a    b   <-- classified as 407 93   a = negative 108 160   b = positive	Attributes: No.1 - plas (0.0986) No.2 - mass (0.0863) No.3 - age (0.0726)	Attributes: No.1 - plas (83) No.2 - mass (70) No.3 - age (31)



### 6.4.2 Constructing Contexts for a Lead Attribute in Pima Indians Diabetes Dataset

After normalizing the original attribute values to the standardized range of [0, 100], these 768 data records are redistributed into four data subsets according to their confusion matrix result categories: 407 in the true negative (TN) subset, 160 in the true positive (TP) subset, 108 in the false positive (FP) subset and 93 in the false negative (FN) subset. Attribute *plas* is selected as the first lead attribute because it is ranked as the most significant attribute by the gain ratio feature evaluator in WEKA.

There are ten key value growth stages based on the sorted deciles and each growth stage value of *plas*, the current lead attribute, together with the other attributes from the same sample instance record, are extracted to prepare for the construction of the incremental and correlational context, as shown in Table 6-3.

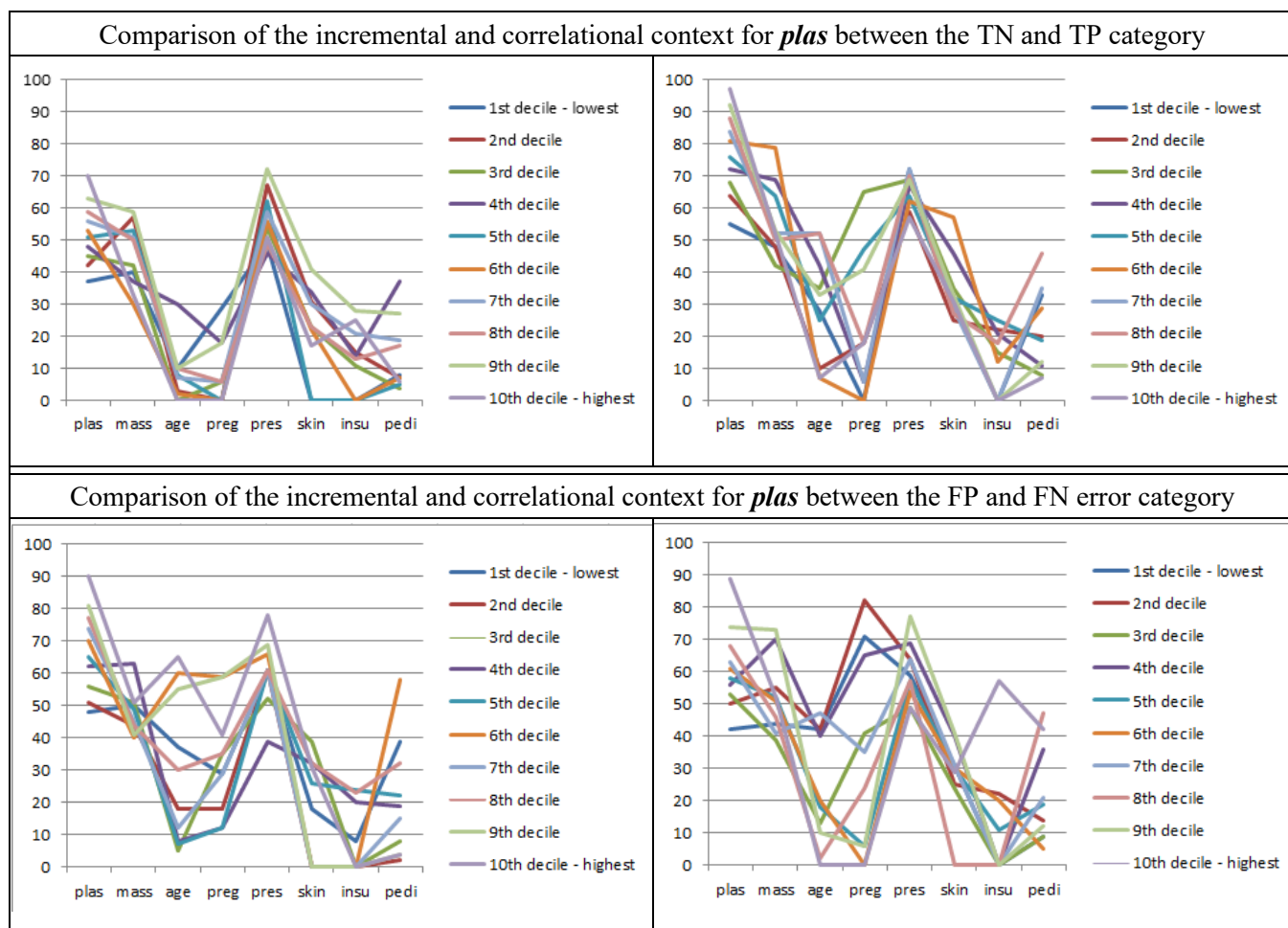
Table 6-3 - Establish the categorical and incremental context for *plas* as the lead attribute

<i>Plas</i> is the lead attribute in this round of incremental sampling by the confusion matrix										
Decile	<b>plas</b>	mass	age	preg	pres	skin	insu	pedi	class	
1 <sup>st</sup> - lowest	<b>37</b>	40	10	29	49	0	0	8	TN	1 <sup>st</sup> - lowest
2 <sup>nd</sup>	<b>42</b>	57	3	0	67	31	15	7	TN	2 <sup>nd</sup>
3 <sup>rd</sup>	<b>45</b>	42	0	6	54	23	11	4	TN	3 <sup>rd</sup>
4 <sup>th</sup>	<b>48</b>	37	30	18	46	34	14	37	TN	4 <sup>th</sup>
5 <sup>th</sup>	<b>51</b>	53	8	0	62	0	0	5	TN	5 <sup>th</sup>
6 <sup>th</sup>	<b>53</b>	30	2	0	56	22	0	7	TN	6 <sup>th</sup>
7 <sup>th</sup>	<b>56</b>	51	7	6	59	30	21	19	TN	7 <sup>th</sup>
8 <sup>th</sup>	<b>59</b>	50	10	6	49	23	13	17	TN	8 <sup>th</sup>
9 <sup>th</sup>	<b>63</b>	59	10	18	72	41	28	27	TN	9 <sup>th</sup>
10 <sup>th</sup> - highest	<b>70</b>	33	0	0	51	17	25	6	TN	10 <sup>th</sup> - highest
Decile	<b>plas</b>	mass	age	preg	pres	skin	insu	pedi	class	
1 <sup>st</sup> - lowest	<b>55</b>	48	28	0	72	30	0	33	TP	1 <sup>st</sup> - lowest
2 <sup>nd</sup>	<b>64</b>	48	10	18	59	25	22	20	TP	2 <sup>nd</sup>
3 <sup>rd</sup>	<b>68</b>	42	35	65	69	35	15	8	TP	3 <sup>rd</sup>
4 <sup>th</sup>	<b>72</b>	69	42	6	67	46	21	11	TP	4 <sup>th</sup>
5 <sup>th</sup>	<b>76</b>	64	25	47	64	32	25	19	TP	5 <sup>th</sup>
6 <sup>th</sup>	<b>81</b>	79	7	0	62	57	12	29	TP	6 <sup>th</sup>
7 <sup>th</sup>	<b>84</b>	52	52	6	72	29	0	35	TP	7 <sup>th</sup>
8 <sup>th</sup>	<b>88</b>	50	52	18	70	27	18	46	TP	8 <sup>th</sup>
9 <sup>th</sup>	<b>92</b>	53	33	41	69	33	0	12	TP	9 <sup>th</sup>
10 <sup>th</sup> - highest	<b>97</b>	52	7	18	57	31	0	7	TP	10 <sup>th</sup> - highest

Decile	plas	mass	age	preg	pres	skin	insu	pedi	class	Decile	plas	mass	age	preg	pres	skin	insu	pedi	class
1 <sup>st</sup> - lowest	48	50	37	29	61	18	8	39	FP	1 <sup>st</sup> - lowest	42	44	42	71	59	31	0	9	FN
2 <sup>nd</sup>	51	44	18	18	61	0	0	2	FP	2 <sup>nd</sup>	50	55	42	82	64	25	22	14	FN
3 <sup>rd</sup>	56	51	5	35	52	39	0	8	FP	3 <sup>rd</sup>	53	39	13	41	49	24	0	9	FN
4 <sup>th</sup>	62	63	8	12	39	32	20	19	FP	4 <sup>th</sup>	56	70	40	65	69	40	0	36	FN
5 <sup>th</sup>	65	49	7	12	61	26	24	22	FP	5 <sup>th</sup>	58	52	18	6	57	30	11	19	FN
6 <sup>th</sup>	70	40	60	59	66	0	0	58	FP	6 <sup>th</sup>	61	51	20	0	54	30	20	5	FN
7 <sup>th</sup>	74	45	12	29	61	0	0	15	FP	7 <sup>th</sup>	63	41	47	35	64	31	0	21	FN
8 <sup>th</sup>	77	44	30	35	61	32	23	32	FP	8 <sup>th</sup>	68	46	2	24	57	0	0	47	FN
9 <sup>th</sup>	81	41	55	59	69	0	0	4	FP	9 <sup>th</sup>	74	73	10	6	77	41	0	12	FN
10 <sup>th</sup> - highest	90	51	65	41	78	31	0	4	FP	10 <sup>th</sup> - highest	89	52	0	0	49	29	57	42	FN

When positioning the lead attribute in the first column and connecting and plot-ting the values across the other attributes from their corresponding records horizontally, the general trends of the line patterns shown in the graphs can represent how these attributes correlate and work together with the lead attribute *plas* along its incremental growth path. Some interesting patterns can be observed amongst these plots, as shown in Table 6-4.

Some obvious value patterns can be observed from these contexts when *plas* is the lead attribute. One distinctive pattern is that the majority of *plas* values in the true negative category are bounded in the middle range between the normalized median value of 37 and 70 and sample values are distributed rather evenly between these four deciles. As the *plas* value increases, the values of the other attributes vary rather uniformly and most are bounded by a narrow value range.

Table 6-4 - Matrix of incremental, correlational and categorical context for *plas*

In contrast, the *plas* values in the true positive category start from a higher point of 55 and scatter in a wider value range up to 97, and the values of the other related attributes, e.g. *mass*, *age* and *pedigree*, also fluctuate in a wider value range with higher starting points. For the two error categories, false positives and false negatives, they share similar bigger and wilder value variation patterns with the true positive category as their *plas* values start from 42 and go up to 90, a wider value range which is more volatile than the true negative.

At a glance, this simple and visual pattern comparison for *plas* can be regarded as a supporting example for the use of contextual analysis by transforming the confusion matrix in the right circumstance. The true negative samples show more uniformity than the true positive samples and the categories of error samples, a distinction which is a common theme for all ten UCI datasets, a kind of close reflection of the Anna Karenina principle.

While it is true to argue that this visual evaluation is too simplistic and does not reveal anything substantial, it illustrates certain contextual differences between result categories and demonstrates one systematic way to construct an indicative contextual model. It seems to be a naïve way to construct contexts, but it is just the beginning.

### **6.4.3 Constructing Contexts for Other Lead Attributes in the Pima Indians Dataset**

When applying this contextual evaluation process to other attributes, it may help to create a fuller picture of the overall contexts for the involved attributes and help prepare the groundwork for a later consolidation of their individual context evaluation.

The attribute *mass* is selected as the next lead attribute in this experiment and its incremental and correlational context are first prepared with their normalized values in a confusion matrix structure as shown in Table 6-5. After this, lines starting from the *mass* column are connected and plotted across values of other attributes of the relevant sample records at each key growth stage and in each category cell to illustrate a three-dimensional contextual model for the attribute *mass*, as shown in Table 6-6.

Table 6-5 - Matrix of incremental, correlational and categorical context for *mass*

<i>Mass</i> is the lead attribute in this round of incremental sampling by the confusion matrix																			
Decile	<i>mass</i>	<i>plas</i>	<i>age</i>	<i>preg</i>	<i>pres</i>	<i>skin</i>	<i>insu</i>	<i>pedi</i>	<i>class</i>	Decile	<i>mass</i>	<i>plas</i>	<i>age</i>	<i>preg</i>	<i>pres</i>	<i>skin</i>	<i>insu</i>	<i>pedi</i>	<i>class</i>
1 <sup>st</sup> - lowest	<b>29</b>	50	15	18	66	11	8	9	TN	1 <sup>st</sup> - lowest	<b>41</b>	46	38	71	51	7	30	36	TP
2 <sup>nd</sup>	<b>34</b>	47	2	6	46	11	0	14	TN	2 <sup>nd</sup>	<b>45</b>	91	65	47	56	36	59	23	TP
3 <sup>rd</sup>	<b>37</b>	56	17	0	53	0	0	25	TN	3 <sup>rd</sup>	<b>47</b>	82	12	18	57	18	12	8	TP
4 <sup>th</sup>	<b>39</b>	67	100	53	61	33	7	16	TN	4 <sup>th</sup>	<b>49</b>	79	17	24	64	0	0	31	TP
5 <sup>th</sup>	<b>42</b>	54	13	6	41	19	0	4	TN	5 <sup>th</sup>	<b>51</b>	95	33	0	85	25	0	15	TP
6 <sup>th</sup>	<b>45</b>	42	0	12	41	23	9	38	TN	6 <sup>th</sup>	<b>54</b>	97	33	41	56	28	0	28	TP
7 <sup>th</sup>	<b>48</b>	46	42	35	51	32	15	0	TN	7 <sup>th</sup>	<b>56</b>	90	2	0	41	36	19	16	TP
8 <sup>th</sup>	<b>51</b>	51	5	0	64	40	11	7	TN	8 <sup>th</sup>	<b>59</b>	79	13	29	69	41	25	14	TP
9 <sup>th</sup>	<b>56</b>	48	10	29	59	33	0	12	TN	9 <sup>th</sup>	<b>64</b>	69	20	0	33	35	20	94	TP
10 <sup>th</sup> - highest	<b>62</b>	43	13	6	54	53	8	36	TN	10 <sup>th</sup> - highest	<b>69</b>	67	42	35	66	37	44	7	TP

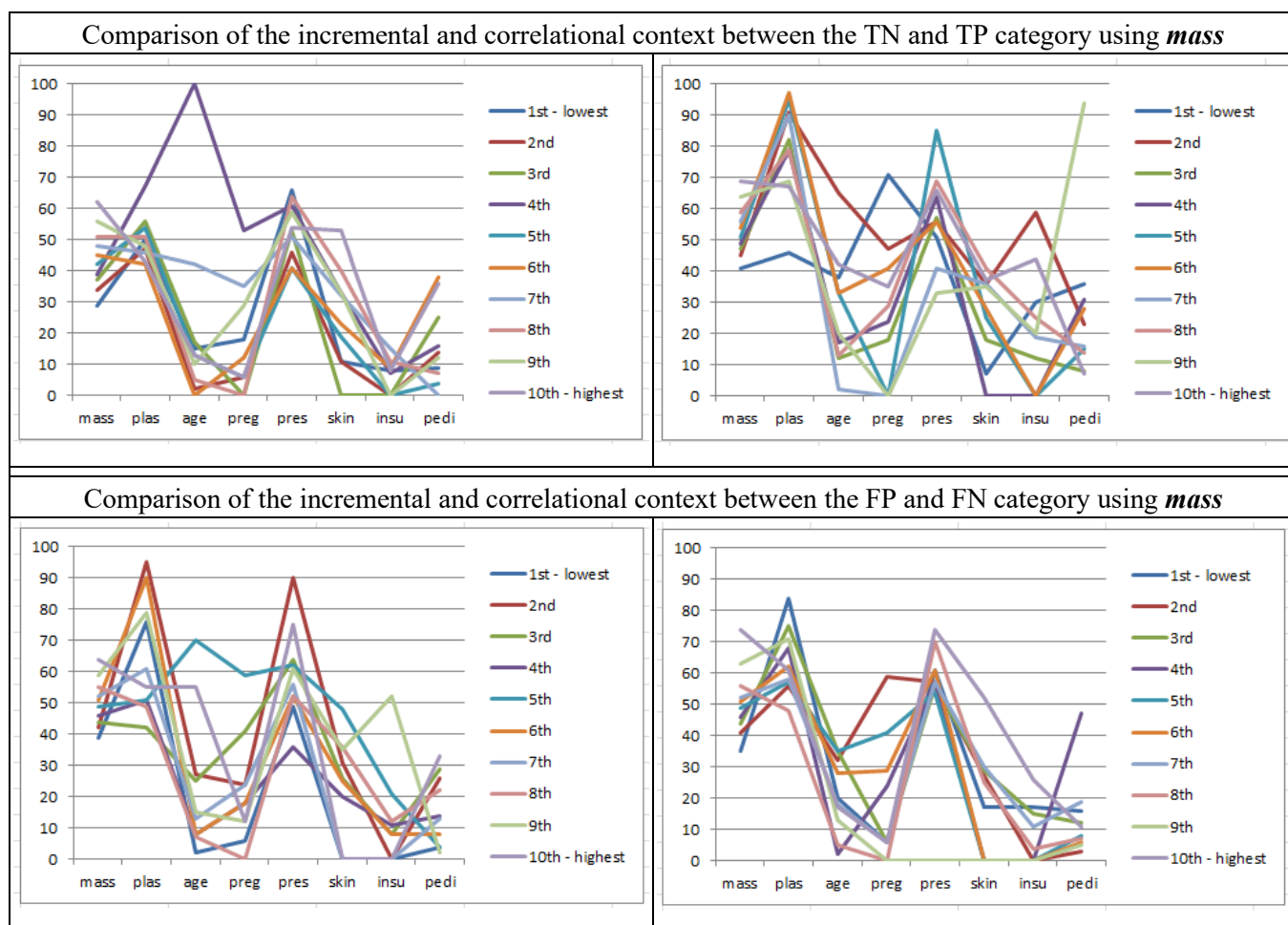
  

Decile	<i>mass</i>	<i>plas</i>	<i>age</i>	<i>preg</i>	<i>pres</i>	<i>skin</i>	<i>insu</i>	<i>pedi</i>	<i>class</i>	Decile	<i>mass</i>	<i>plas</i>	<i>age</i>	<i>preg</i>	<i>pres</i>	<i>skin</i>	<i>insu</i>	<i>pedi</i>	<i>class</i>
1 <sup>st</sup> - lowest	<b>39</b>	76	2	6	49	0	0	4	FP	1 <sup>st</sup> - lowest	<b>35</b>	84	20	6	61	17	17	16	FN
2 <sup>nd</sup>	<b>42</b>	95	27	24	90	31	0	26	FP	2 <sup>nd</sup>	<b>41</b>	56	32	59	57	27	0	3	FN
3 <sup>rd</sup>	<b>44</b>	42	25	41	64	26	8	29	FP	3 <sup>rd</sup>	<b>44</b>	75	35	6	56	29	15	12	FN
4 <sup>th</sup>	<b>46</b>	51	8	18	36	20	11	14	FP	4 <sup>th</sup>	<b>46</b>	68	2	24	57	0	0	47	FN
5 <sup>th</sup>	<b>49</b>	51	70	59	62	48	21	4	FP	5 <sup>th</sup>	<b>49</b>	57	35	41	54	0	0	8	FN
6 <sup>th</sup>	<b>51</b>	90	8	18	52	25	8	8	FP	6 <sup>th</sup>	<b>51</b>	62	28	29	61	0	0	6	FN
7 <sup>th</sup>	<b>52</b>	61	13	24	56	0	0	13	FP	7 <sup>th</sup>	<b>52</b>	58	18	6	57	30	11	19	FN
8 <sup>th</sup>	<b>55</b>	49	7	0	52	36	12	22	FP	8 <sup>th</sup>	<b>56</b>	48	5	0	70	25	4	7	FN
9 <sup>th</sup>	<b>59</b>	79	15	12	61	35	52	2	FP	9 <sup>th</sup>	<b>63</b>	71	13	0	0	0	0	5	FN
10 <sup>th</sup> - highest	<b>64</b>	55	55	12	75	0	0	33	FP	10 <sup>th</sup> - highest	<b>74</b>	61	17	6	74	52	26	11	FN

One distinctive pattern when *mass* is the lead attribute can be observed in Table 6-5 and 6-6, that is, for most true negative samples, the values of all attributes are within their lower six deciles and have small variation margins along the value growth path of *mass*, and the false negative samples have similar patterns except for some *plas* and *pres* values which are above the sixth decile. In contrast, the true positive and false positive samples

have many values above the sixth decile and have bigger and wilder variation margins along the value growth path of *mass*.

Table 6-6 – Plotting the matrix of incremental, correlational and categorical context



Another level of contextual analysis is the comparison between different lead attributes, for example, the contextual model for *plas* as the lead attribute in Table 6-4 and *mass* as the lead attribute in Table 6-6. To be more specific, the context comparison of true negative samples shows that when *mass* is the lead, it has the incremental range [29~62] and its associated *plas* attribute has the value range [42~67]; and when *plas* is the lead, it has the incremental range [37~70] and its associated *mass* has the value range

[33~59], a much smaller variation margin compared to the other attributes. This close-coupled pattern can be an indication of a close association between *plas* and mass, as proved by the gain ratio evaluator in WEKA, shown in Table 6-2.

Context construction and evaluation of the other attributes of this Pima Indians diabetes dataset also present similar patterns, and most of them produce similar results which indicates that the true negatives and healthy samples share more consistent and predictable value patterns within their incremental and correlational context; whereas, true positives and errors, meaning the unhealthy and error-prone samples, have more inconsistent and unpredictable value patterns.

This again demonstrates how contextual analysis may be useful in connecting and comparing attributes and patterns for a better understanding.

#### **6.4.4 Constructing Contexts for Other Datasets**

Other interesting patterns have been also observed in experiments with nine other UCI datasets. One such example is the experiment with *length* as the lead attribute in the Page Blocks dataset. After the completion of an initial classification process and all involved attribute values are normalized, an attribute rotation list for key attribute selection and correlation illustration can be determined, for example, a highly ranked significant attribute by the gain ratio evaluator or a highly ranked error-sensitive attribute by the attribute-error counter. The most significant attribute *length* by the gain ratio is selected as the starter in the key attribute selection list in this experiment, as shown in Table 6-7.

Table 6-7 - Classification result for Page Blocks dataset and options for lead attributes

Correctly Classified: 5321 <b>97.22%</b> Incorrectly Classified: 152 2.78%  === Confusion Matrix === a  b  <-- classified as 4844  69     a = text 83 477    b = non_text	Top-3 attributes ranked by gain ratio evaluator  Attributes: No.1 - length (0.2993) No.2 - height (0.1494) No.3 - mean_tr (0.143)	Top-3 attributes ranked by attribute-error counter  Attributes: No.1 - mean_tr (89) No.2 - p_black (43) No.3 - eccen (29)
---	---	---

Based on the initial classification result and its confusion matrix, data samples are redistributed to four data subsets according to their confusion matrix result categories, and ten key growth stages of the *length* attribute value can be established based on the median value their deciles in the corresponding sample subset, as shown in Table 6-8.

One form of interrelationship between attributes during the value growth of *length* can subsequently be illustrated by plotting each key growth stage value of *length* and across the other attributes of the same selected sample record and the fluctuations in the general trends of the line patterns can be an indication of how other attributes are influencing *length* actively or how they are impacted by *length* reflexively in a correlated way, as shown in Table 6-1.

Such contextual comparison of value patterns from a multi-dimensional perspective is illustrated again in Table 6-9 for easier comparison and cross-matching with Table 6-8, and one obvious value pattern which can be observed from the above matrix of contexts is the contextual difference between the true positive (text) samples and other categories in the form of line pattern formation.



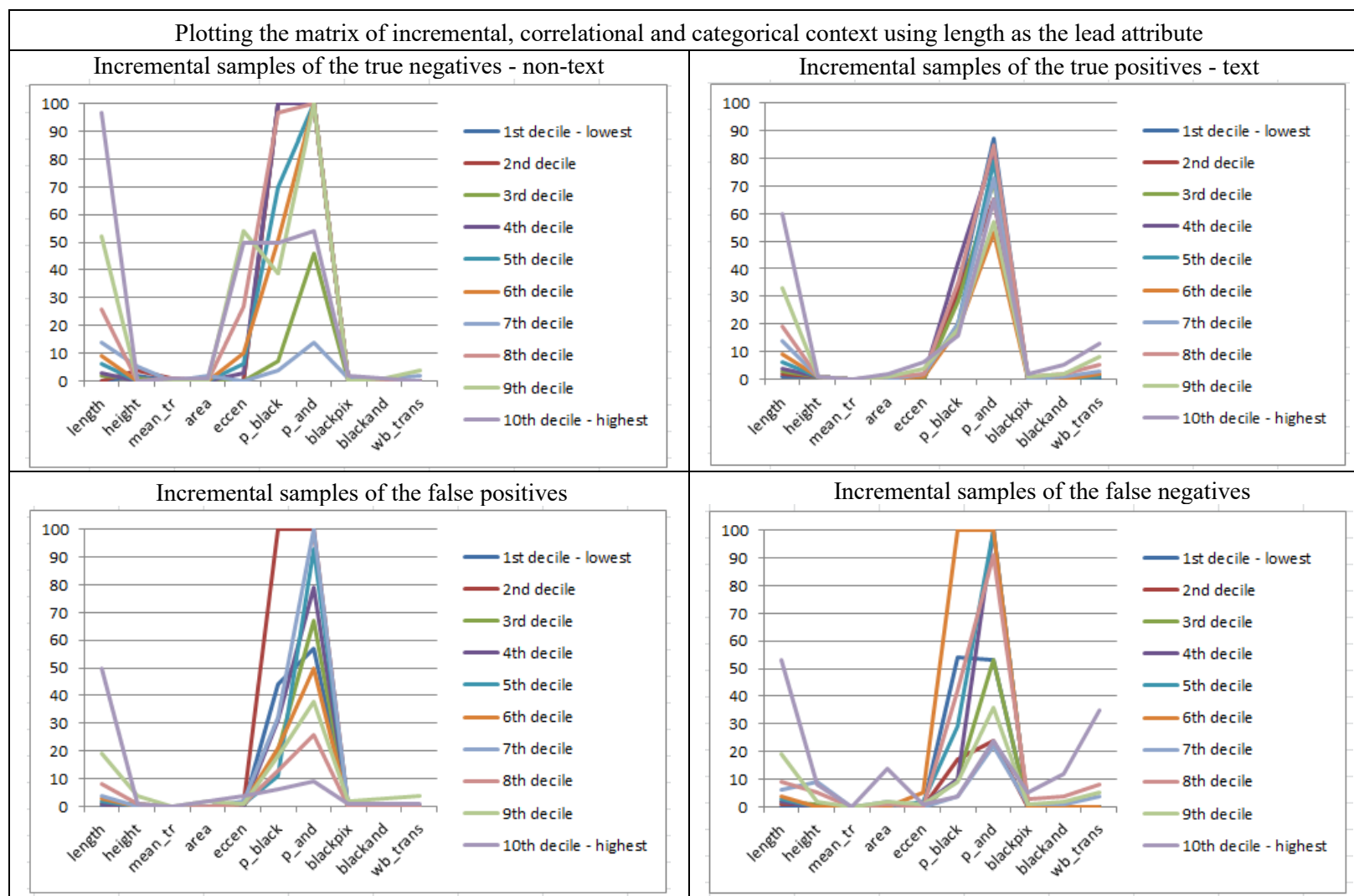
Table 6-8 - Establishing categorical and incremental context for *length* as the lead attribute

<i>Length</i> is the lead attribute in this round of incremental sampling by the confusion matrix											
Decile by length	length	height	mean_tr	area	eccen	p_black	p_and	blackpix	blackand	wb_trans	class
1st decile - lowest	0	2	0	0	0	100	100	0	0	0	TN
2nd decile	0	4	1	0	0	100	100	0	0	0	TN
3rd decile	2	1	0	0	0	7	46	0	0	0	TN
4th decile	3	0	0	0	3	100	100	0	0	0	TN
5th decile	6	0	0	0	6	70	100	0	0	0	TN
6th decile	9	0	0	0	10	51	100	0	0	0	TN
7th decile	14	5	0	2	0	4	14	1	1	2	TN
8th decile	26	0	1	0	27	97	100	0	0	0	TN
9th decile	52	0	0	0	54	39	100	0	1	4	TN
10th decile - highest	97	0	1	1	50	50	54	2	1	0	TN

Decile by length	length	height	mean_tr	area	eccen	p_black	p_and	blackpix	blackand	wb_trans	class
1st decile - lowest	1	1	0	0	0	31	87	0	0	0	TP
2nd decile	2	1	0	0	0	32	65	0	0	1	TP
3rd decile	3	1	0	0	0	28	72	0	0	1	TP
4th decile	4	1	0	0	1	42	82	0	0	1	TP
5th decile	6	1	0	0	1	20	79	0	0	1	TP
6th decile	9	1	0	0	1	18	53	0	0	2	TP
7th decile	14	1	0	0	2	20	73	0	1	3	TP
8th decile	19	1	0	1	2	35	85	1	2	5	TP
9th decile	33	1	0	1	4	18	57	1	2	8	TP
10th decile - highest	60	1	0	2	6	16	65	2	5	13	TP

Decile by length	length	height	mean_tr	area	eccen	p_black	p_and	blackpix	blackand	wb_trans	class
1st decile - lowest	0	1	0	0	0	54	53	0	0	0	FP
2nd decile	1	1	0	0	0	17	24	0	0	0	FP
3rd decile	2	1	0	0	0	10	53	0	0	0	FP
4th decile	2	0	0	0	1	10	100	0	0	0	FP
5th decile	3	0	0	0	2	29	100	0	0	0	FP
6th decile	4	0	0	0	5	100	100	0	0	0	FP
7th decile	6	9	0	2	0	4	22	1	1	4	FP
8th decile	9	5	0	1	0	42	91	3	4	8	FP
9th decile	19	2	0	2	1	9	36	1	2	5	FP
10th decile - highest	53	8	0	14	1	4	24	5	12	35	FP

Decile by plas	length	height	mean_tr	area	eccen	p_black	p_and	blackpix	blackand	wb_trans	class
1st decile - lowest	0	0	0	0	0	44	57	0	0	0	FN
2nd decile	1	0	0	0	1	100	100	0	0	0	FN
3rd decile	1	1	0	0	0	20	67	0	0	0	FN
4th decile	1	0	0	0	0	31	79	0	0	0	FN
5th decile	2	0	0	0	1	11	93	0	0	0	FN
6th decile	3	0	0	0	1	21	50	0	0	0	FN
7th decile	4	0	0	0	2	32	100	0	0	0	FN
8th decile	8	1	0	0	1	13	26	0	0	0	FN
9th decile	19	4	0	2	1	18	38	2	3	4	FN
10th decile - highest	50	1	0	2	4	6	9	1	1	1	FN

Table 6-9 - Matrix of incremental, correlational and categorical context for *length* as the lead attribute

At a glance, the correlational plots across attributes based on the incremental path of lead attribute *length* are shown to rise and fall in a synchronized and closely bound uniformity for the true positive (text type) data samples; as for the true negative (non-text type) data samples and the samples that have been misclassified, the correlational plots across attributes based on the incremental path of lead attribute *length* do not show such synchronized variation. The contrast between the harmonious plotting patterns of the true positive (text type) data samples and the unsynchronized crisscrossing patterns of the rest is rather palpable, and further analysis and discussion of such discrepancy between class labels and errors are provided in the next section.

As shown in the contrasting line patterns between the true positive (text) samples and the other three categories, the true positive (text) samples show close-coupled lines between *length* and the other attributes, especially the consistent and narrow value range for the *p\_black* (percentage of black pixels within the block) and the *p\_and* (percentage of black pixels after the application of the run length smoothing algorithm) attribute in high deciles, which is a strong indication of black and regular text blocks, which in turn, becomes a sign of correlation between the true positive (text) category and the long and black regular text blocks.

On the other hand, the true negative (non-text) and error samples show widely fluctuating line patterns, especially the large value margin in attributes *p\_black* and *p\_and*, indicating a large variation in color pigment and text font-size, therefore they are more likely correlated with colorful and rich format graphics or hyper markup text blocks which

can be a source of confusion and misclassification for the classifier to determine if a page block is really in the format of text or not.

Another interesting contextual value pattern can be observed in another dataset, the Wisconsin cancer dataset, when the attribute *Normal\_Nucleoli*, the most significant attribute ranked by gain ratio is selected as the lead attribute for context construction and analysis.

Based on the initial classification result, the incremental samples for the selected lead attribute *Normal\_Nucleoli* can be prepared in ten incremental growth stages and categorized under the resulting confusion matrix structure, as shown in Table 6-10. To help visualize the correlation between attributes for the ten growth stages and between the result categories, lines are plotted from attribute *Normal\_Nucleoli* at each growth stage and connected across the other involved attributes to illustrate the potential impact of or association with the lead attribute *Normal\_Nucleoli*, as shown in Table 6-11.

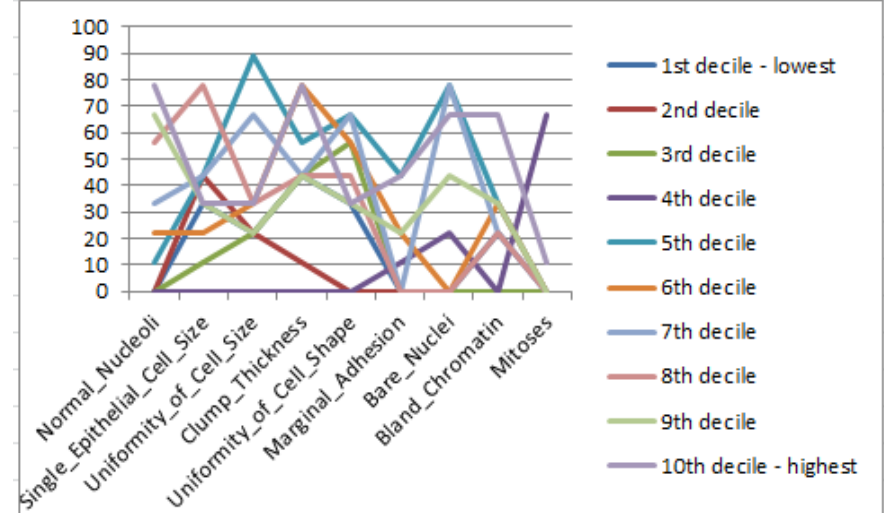
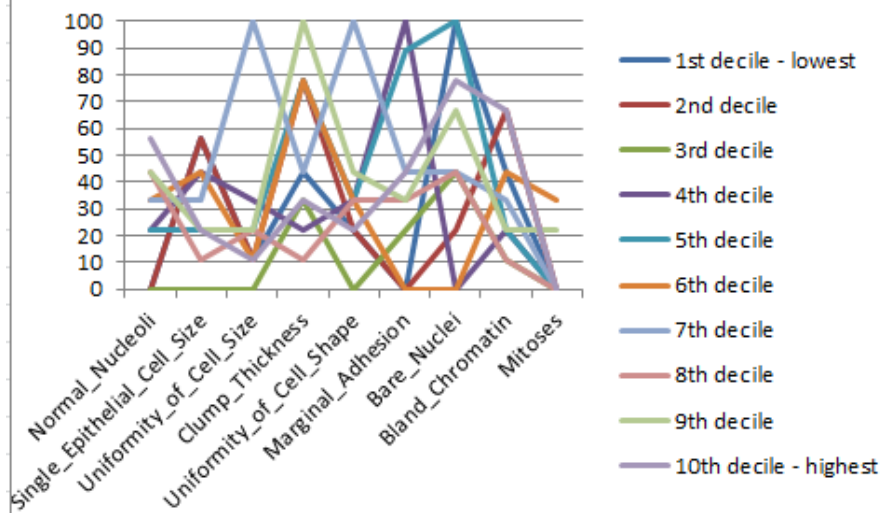
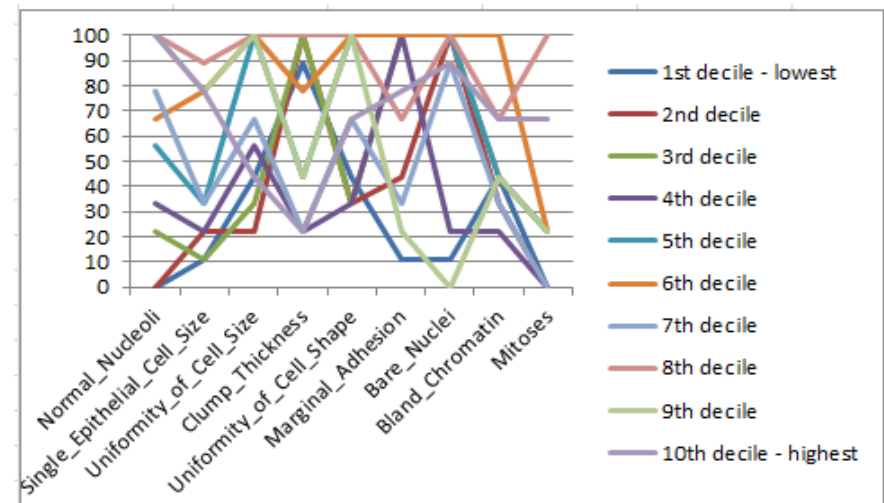
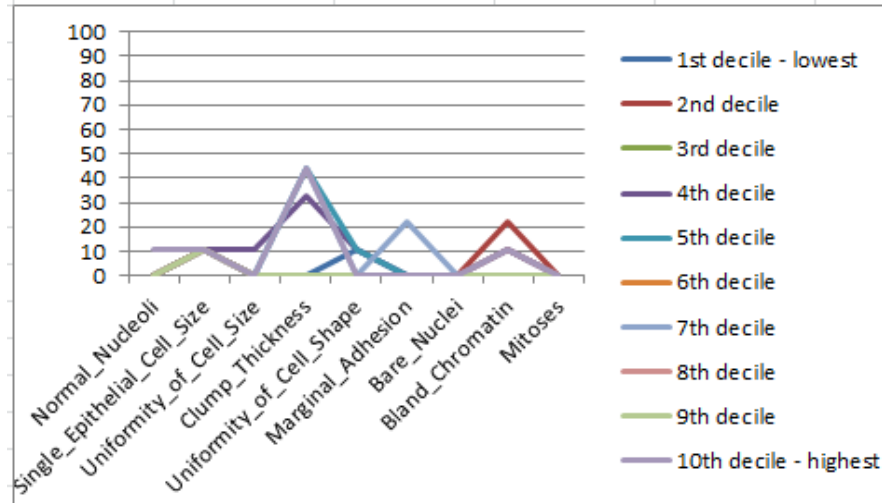
This contextual model based on *Normal\_Nucleoli* as the lead attribute in the Wisconsin cancer dataset again clearly reflects the Anna Karenina principle, that is, for the samples in the true negative (TN) healthy category as shown in the top left diagram of Table 6-11, their *Normal\_Nucleoli* values are consistently low between 0 and 10 in all ten incremental growth stages, and so are most of the values from the other related attributes, which mostly remained under the value of 20, so this is a kind of “Happy families are all alike”.

Table 6-10 - Establishing categorical and incremental context for *Normal\_Nucleoli* as the lead attribute

<i>Normal_Nucleoli</i> is the lead attribute in this round of incremental sampling by the confusion matrix																					
Decile by Normal_Nucleoli	Normal Nucleoli	Single_Epithelial Cell_Size	Uniformity of Cell_Size	Clump Thickness	Uniformity_of Cell_Shape	Marginal Adhesion	Bare Nuclei	Bland Chromatin	Mitoses	Class	Decile by Normal_Nucleoli	Normal Nucleoli	Single_Epithelial Cell_Size	Uniformity of Cell_Size	Clump Thickness	Uniformity_of Cell_Shape	Marginal Adhesion	Bare Nuclei	Bland Chromatin	Mitoses	Class
1st decile - lowest	0	11	0	0	11	0	0	11	0	TN	1st decile - lowest	0	11	44	89	44	11	11	44	0	TP
2nd decile	0	11	0	0	0	0	0	22	0	TN	2nd decile	0	22	22	100	33	44	100	33	0	TP
3rd decile	0	11	0	0	0	0	0	0	0	TN	3rd decile	22	11	33	100	33	100	100	44	22	TP
4th decile	0	11	11	33	11	0	0	11	0	TN	4th decile	33	22	56	22	33	100	22	22	0	TP
5th decile	0	11	0	44	11	0	0	0	0	TN	5th decile	56	33	100	44	100	100	100	44	22	TP
6th decile	0	11	0	0	0	0	0	0	0	TN	6th decile	67	78	100	78	100	100	100	100	22	TP
7th decile	0	11	0	44	0	22	0	0	0	TN	7th decile	78	33	67	22	67	33	89	33	0	TP
8th decile	0	11	0	0	0	0	0	11	0	TN	8th decile	100	89	100	100	100	67	100	67	100	TP
9th decile	0	11	0	0	0	0	0	0	0	TN	9th decile	100	78	100	44	100	22	0	44	22	TP
10th decile - highest	11	11	0	44	0	0	0	11	0	TN	10th decile - highest	100	78	44	22	67	78	89	67	67	TP
Decile by Normal_Nucleoli	Normal Nucleoli	Single_Epithelial Cell_Size	Uniformity of Cell_Size	Clump Thickness	Uniformity_of Cell_Shape	Marginal Adhesion	Bare Nuclei	Bland Chromatin	Mitoses	Class	Decile by Normal_Nucleoli	Normal Nucleoli	Single_Epithelial Cell_Size	Uniformity of Cell_Size	Clump Thickness	Uniformity_of Cell_Shape	Marginal Adhesion	Bare Nuclei	Bland Chromatin	Mitoses	Class
1st decile - lowest	0	56	11	44	22	0	100	44	0	FP	1st decile - lowest	0	33	22	44	33	0	0	22	0	FN
2nd decile	0	56	11	78	22	0	22	67	0	FP	2nd decile	0	44	22	11	0	0	0	0	0	FN
3rd decile	0	0	0	33	0	22	44	11	0	FP	3rd decile	0	11	22	44	56	0	0	0	0	FN
4th decile	22	44	33	22	33	100	0	22	0	FP	4th decile	0	0	0	0	0	11	22	0	67	FN
5th decile	22	22	22	78	33	89	100	22	0	FP	5th decile	11	44	89	56	67	44	78	33	0	FN
6th decile	33	44	11	78	33	0	0	44	33	FP	6th decile	22	22	33	78	56	22	0	33	0	FN
7th decile	33	33	100	44	100	44	44	33	0	FP	7th decile	33	44	67	44	67	0	78	22	0	FN
8th decile	44	11	22	11	33	33	44	11	0	FP	8th decile	56	78	33	44	44	0	0	22	0	FN
9th decile	44	22	22	100	44	33	67	22	22	FP	9th decile	67	33	22	44	33	22	44	33	0	FN
10th decile - highest	56	22	11	33	22	44	78	67	0	FP	10th decile - highest	78	33	33	78	33	44	67	67	11	FN

Table 6-11 - Matrix of incremental, correlational and categorical context for *Normal\_Nucleoli*

*Normal\_Nucleoli* is the lead attribute in this round of incremental sampling by the confusion matrix



In contrast, in the top right diagram and the two bottom ones of Table 6-11 that represent the samples in the other three categories, the true positive (TP), false positive (FP) and false negative (FN), their *Normal\_Nucleoli* values grow steadily from 0 to 100, and the values of the other related attributes fluctuate widely, so this is kind of “every unhappy family is unhappy in its own way”.

## 6.5 Discussion of Potential Issues

While such distinctive contextual value patterns may look interesting, some may be explained by statistical or more general theories like the Anna Karenina principle, a term based on Leo Tolstoy's novel Anna Karenina and is partially about the contrast and transition between success from harmonious unity and failure due to chaos (more detail on the Anna Karenina principle is discussed in Section 6.6.2). Contextual value patterns can be tested and verified by renowned algorithms, such as the decision tree with information gain or the gain ratio method. In reality, they may not mean anything other than diagrams with some distinctive patterns for the selected attributes and selected datasets. This is similar to enhancing classification accuracy by algorithm development but with no intention to understand the data elements better when the research goal is data classification and analysis.

It is believed that inviting domain experts with in-depth field knowledge to join the exploration of contexts may be a more productive way to conduct data mining research on domain-specific topics. Having the right context and interpretation can be more important than making a quick claim. Simplicity can sometimes mean naivety and liability. Context construction by incremental sampling, cross-attribute plotting and matrix category cross-matching may sound trivial, and its representation of incremental, correlational and

categorical context may look inconsequential. On the other hand, the focus of this study is not to make a breakthrough in algorithm performance and complexity, rather it is about sorting and connecting data elements to gain a better understanding of data within a context and to start such contextual analysis with simplicity and practicality. Further improvements with the inclusion of ROC curve analysis, F-score and other techniques can be considered next after the conclusion of current study.

A more specific issue concerns the suitability and adequacy of using ten sorted deciles to represent ten value growth stages of a lead attribute, although this is not a rigid way of data sampling and its value growth representation can be over-optimistic. On the other hand, as stated by Cochran [Coc97], “systematic sampling provides a kind of stratification with equal sampling fractions ... (and) systematic samples are convenient to draw and execute” (Cochran, 1997, p.226). Based on this argument, using deciles in an overview of value growth in terms of incremental context can be sufficient as a starting point.

The use of the median value of a decile to represent its corresponding key growth stage value can be another concern. Other options to consider include using the average value of the selected lead attribute in each decile, the value of mode in each decile, and the weighted average value in each decile, etc. The main justification for using the median value in this experiment is its ease of retrieving all the other related attribute values for correlation analysis once the sample record is identified with its selected lead attribute which has the median value of the corresponding decile, therefore the correlational context constructed for each growth stage in this case is based on a complete and realistic sample record as a whole, not through the repeated computation of individual average, or mean, or



mode, or weighted average for each other attribute involved in the corresponding decile to represent an artificial correlational context.

Another serious issue is the use of the min-max normalization method. Its application and result can be impacted by distorted value conversion and outlier values at both min and max ends, and the use of a median record from each decile is one attempt to reduce such impact. In the next stage, Z-score-based normalization can be an alternative, but it requires further consideration and modification to address the issues, such as one standardized scale may not be suitable for all involved attributes with different measures.

There is no doubt that there is a long list of other problems and issues in relation to this study, and these problems and issues can also become the contexts of further discussion and improvement. On the other hand, one key objective of this context construction idea is to identify contextual information patterns in relation to specific value patterns, to raise the awareness of the stakeholders and to help explore a further understanding and discovery of data.

While this study has only constructed and outlined three dimensions in the current context model, incremental, correlation and categorical, this context construction process can later be expanded to include other dimensions, such as a temporal and geographical context, to provide a bigger picture with more correlated contexts, internally and externally. Such an expansion of context analysis may require more attention and effort to explore and examine when data comprehension and analysis is the priority rather than classification accuracy.

## 6.6 Related Work

The usual focus of data mining and classification research is on performance enhancement in terms of theories and algorithms. Using a decision tree classification framework as an example, C4.5 [Qui93] and CART [BFO84] are two early benchmark models, and numerous enhancements on various model components have been developed over the years. Individual trees have been “growing” and “spreading” into ensemble models such as random forest, Ada trees, probabilistic boosting-trees and a combined Bayesian tree model using innovative sampling and weighing techniques [GE04] [Tu05] [MCS11] [RM14]. Amidst the large volume of literature reporting the rapid development and progress on theories and algorithms that are competing for better performance, there appears to be much less study and development on how such progress and results can be utilized to improve the understanding of data elements and their correlation with the result categories and errors in a systematic and integrated way. This study attempts to address this.

### 6.6.1 Study of Confusion Matrix and Contextual Analysis

A confusion matrix is a form of contingency table or frequency table. While a frequency table is used to group and tally specified objects or events by category, a confusion matrix in the field of data mining and classification is used to tally and tabulate the predicted result by class group in columns, and the actual result of each class group in rows, as shown in Table 6-12, and the organization of such columns and rows can be interchangeable [KP98] [Faw06].

Table 6-12 –Confusion matrix for two-class classification

		Predicted result	
		Negative	Positive
Actual result	Negative	a	b
	Positive	c	d

This confusion table can be interpreted as follows:

- a is the number of true negative instances;
- b is the number of false positive instances;
- c is the number of false negative instances;
- d is the number of true positive instances.

A confusion matrix is a simple and effective way to summarize and compare classification results and has also recently been used in attribute selection model development and error detection [VRR11] [PBD10], but it is predominately being used to highlight the difference in result categories rather than being used to compare the paths leading to the different decisions and the context for such difference [HGC16].

The new idea of utilizing the confusion matrix as an attribute selection mode is to introduce a disagreement score [VRR11] for attribute ranking and selection. To define the disagreement as 1 if the number value of b or c is 0 is a strong indication that there is less chance of error and higher discrimination power at least in one result class group. To define the score as 0 if b and c are the same is a sign of a higher chance of error and less discrimination power. We find the optimum balance between a disagreement score (D) and a complementary combination of attribute subset by making use of the value of b and c from each attribute in the following formula:

$$D = \begin{cases} 0 & \text{if } b = c = 0; \\ \frac{|b-c|}{\max\{b,c\}} & \text{otherwise.} \end{cases} \quad (6.2)$$

After the attributes are processed and ranked by their individual disagreement score  $D$ , the highly ranked attributes with misclassification errors in different result class groups are selected to form a series of complementary attribute subsets as optimization candidates for further examination, and each attribute within a subset is supposed to complement each other in misclassifying a different class group so the newly formed subset as a whole can deliver better performance with higher accuracy when adopted in another round of classification process.

While this attribute selection by confusion matrix approach is indeed an interesting one, it is not considered an effective solution when dealing with large amounts of data with many attributes and many class groups, which may lead to a large number of complementary attribute combinations generated as candidate subsets that are subsequently required for testing, comparison and impact analysis.

There has also been criticism of the confusion matrix over its inadequacy in dealing with imbalanced class labels and cost sensitive classification issues. For example, in the case of medical diagnostic data classification, if errors such as false negatives are fewer than false positives, the cost can be far more expensive and the associated risk can be far more serious; therefore, more advanced measures for performance results, such as the receiver operating characteristic (ROC) curve and F-score should be considered more favorably [PFK98] [SJS06] [Faw06]. While agreeing with the criticism, this study takes a

pragmatic approach in adopting the confusion matrix because of its simplicity and its capability of accommodating and visualizing more forms of context in a single place.

One apparent path which may influence the classification process and its confusion matrix outcome could be the value growth path of certain key attributes. In dealing with a large amount of data, one way to observe such value variation and growth trend is through sampling. The construction method of incremental context in this study is a kind of data sampling. Of the many well-researched data sampling techniques [Loh09] [PHG15], this study adopts the systematic sampling approach with regular intervals. Despite being a simplistic approach and only working well with regular values [Coc97], and recognizing the need to reduce data dimensionality, redundancy and noise [HG03] [LY05] [DX13], systematic sampling can still be argued, though vaguely, to be an effective tool to establish the value incremental context in this pilot study.

The value growth of an attribute may influence or be impacted by the value variation of other attributes, but the degree of influence may depend on specific attributes, therefore attribute prioritization and selection can play an important role in data classification when dealing with a large amount of data with many attributes. Research on ways to reduce data dimensionality, redundancy and noise is ongoing [HG03] [LY05], and examples of new approaches include the fuzzy rough set-based information gain model and the error-sensitive attribute selection model based on attribute error counts [DX13] [WZ13].

External contexts can influence the data as a kind of leading cause and driving force, one example reported in Phi1's research on the commonly used Pima Indian diabetes data [Phi12]. It reports that the inadequately documented geographical and historical

reasons, together with economic and social factors, played a more significant role in causing diabetes amongst the Pima Indians than the well-documented and highly-ranked therapeutic attributes, such as diabetes pedigree, body mass index, blood pressure and plasma glucose level. Phihl's research has particularly singled out the loss of access to water from the Gila River in the 1890s as the key trigger point and the historical and geographical context and cause for the diabetes epidemic amongst the Pima Indians in the Arizona region.

Other research on the Pima Indian diabetes data has also studied potential causes and impact from historical, , environmental, social and economic contexts [Nar97] [SBR06] [VWN05] in a more general sense, and some forms of internal and implicit contexts in terms of body metabolism and energy circulation are discussed in various alternative medical research reports [Cov01] [WWC13].

While the identification and analysis of external contexts is important, they are not included in the current scope of this study.

### **6.6.2 The Adoption of Anna Karenina Principle in Other Studies**

The reference to the Anna Karenina principle in this contextual analysis seemed farfetched in the beginning. The original novel "Anna Karenina" by Leo Tolstoy in 1877 was about the struggles between the pursuit of freedom of passion and individual happiness and the resistance of family traditions and society expectations, and there is not really a scientific law or rule named the Anna Karenina principle. So, within a narrow and literal context, the reference to the Anna Karenina principle may not make sense.

On the other hand, if the essence of the Anna Karenina principle can be interpreted as the contrast between the harmonization and the aberration of various related factors of a task and in a wider context of nature and science observation, then such a reference subsequently makes more sense and has been adopted and examined by a number of studies over the years. For example, the Anna Karenina principle was used by Jared Diamond to justify why only a few societies or civilizations, i.e. Eurasian and North African civilizations, were able to survive, prosper, and conquer while many others failed during the brutal and destitute ancient history for thousands of years [Dia98], as summarized in Wikipedia about this book in regard to Diamond's theory - "that Eurasian civilization is not so much a product of ingenuity, but of opportunity and necessity. That is, civilization is not created out of superior intelligence, but is the result of a chain of developments, each made possible by certain preconditions." ("Guns, Germs, and Steel", n.d.). In suggesting some of the favorable preconditions such as vast fertile land, suitable plants for a reliable food supply and division of labor into specialized skills like literacy and mining, Diamond also used the Anna Karenina principle to explain why only a small number of animal species had been domesticated for the purpose of food supply, field work and transport, saying "many promising species have just one of several significant difficulties that prevent domestication", therefore, "domesticable animals are all alike; every undomesticable animal is undomesticable in its own way." (Diamond, 1998, p.157)

Another recent example of applying this principle is the research into how the historical and statistical analysis on key paper publications on the plate tectonics concept can help identify the "specific prerequisites" and functional details necessary for the

“breakthroughs” and “revolutions” in “geoscientific thinking” [MB13], and how this concept development process reflects the Anna Karenina principle very well by showing its progress stalled time and again when various key conditions and factors were missing or overlooked, just like the “each unhappy family is unhappy in its own way”; and how the consensus on “the paradigm shift took from fixism to plate tectonics” was reached “which was finally accepted by the scientific community in the geosciences” (Marx and Bornmann, 2013, p.17) in the science community after more than 50 years of study and debate by so many researchers, just like “All happy families are alike” after all.

## 6.7 Summary

This suggestion to transform a confusion matrix into a matrix of incremental, correlational and categorical context to analyze data and classification results may seem far-fetched, but initial experiments reveal some supportive signs which reflect well the Anna Karenina principle even though in an ironic sense. The closely uniform and harmoniously synchronized value patterns in the true negative (TN) and healthy data samples in the diagnostic datasets, such as the Pima Indian diabetes and Wisconsin cancer datasets, and the similarly uniform and synchronized value patterns in the true positive (TP) and text-type samples of the Page Blocks dataset, pose a form of context that the suggested attributes would need to maintain in terms of a certain level of coordination and adapt to certain rigid conditions in order to stay healthy and “stick to one’s word” (in terms of text) just like doing all the right things mutually required for a happy family. On the other hand, if a value of any related attribute is out of a certain safety zone, then this would potentially lead to the opposite result or misclassification error due to the disruption to the biologically or naturally



balanced interrelation and inter-reaction between attributes and the subsequent loss of balance and lack of harmony between interrelated attributes. Contextual analysis aims to identify such explicit and implicit contexts to help improve the understanding of the data, the result and the errors.

Furthermore, the key message of this study is, while the pursuit of performance and accuracy enhancement is inspirational and admirable, a little detour to explore inside the data and classification results for context construction and analysis may also be meaningful and beneficial in identifying internal and implicit contexts, and may lead to further understanding and knowledge discovery of the data, the results, the errors and the related factors by extracting and expanding, integrating and correlating value patterns, and reviewing and examining the patterns and findings from multiple perspectives and within contexts.

This context construction process can be considered an attempt to define and construct internal context from within the data and its classification result, but it is only a starting point. The next stage of development may include features such as a zooming and binning capability into the process, so context construction can be localized into problematic value patterns across multiple attributes to have a more specific and relevant context for a better understanding.

## CHAPTER 7

### **Conclusion and Future Work**

Research on data mining and data classification has primarily focused on the components and techniques of the specific mining and classification process, such as theory and algorithm development for performance enhancement. The role of a dataset is usually auxiliary and is mainly used for testing, to be used in experiments to prove a successful data mining and classification case after which there is no more further review or examination into the data values from the standpoint of the dataset to develop more connections and comprehension.

In recognizing this issue, and as an attempt to explore and examine the further correlation between classification results and value patterns with the goal of gaining more understanding about different aspects of the data, a research journey has been conducted to explore error-sensitive value patterns in association with classification processes from the perspective of value ambiguity, risk and error-sensitivity, progressing to transparency and contextuality.

The major components of this research journey have been reported in this thesis and the key findings can be summarized as follows.

#### **7.1 Summary of the Journey and Findings**

This research journey started with a study into the ambiguous value range in terms of binary data classification as discussed in Chapter 3 to measure the width and depth of the grey

area between the positive and negative sample values by introducing the error-sensitive attribute evaluation process to provide some form of context in understanding how one attribute's value transition may be related to a potential value range threshold with substantial influence for the change from negative to positive or vice versa, which may also lead to a higher chance of misclassification when the attribute value of a sample is within such an ambiguous and error-sensitive range.

Initial experiments produced some encouraging results and findings to support this idea and its proposed error-sensitive evaluation process when testing with UCI datasets. When comparing this evaluation process against the renowned gain ratio and information gain algorithm, the majority of the most at-risk and error-sensitive attributes identified are also the most significant attributes ranked by the gain ratio and information gain algorithm. When re-classification is conducted after filtering out some of the most error-sensitive attributes as an investigative form of error reduction, three datasets showed increased accuracy in various ways, and even though two datasets did not show improved results, there seems to be valid reasons to explain such unsupportive results, so in a way this can also be considered a justifiable outcome.

The next step of the journey moved on to a more detailed and specific level study of risky and error-sensitive patterns, as discussed in Chapter 4, and the study developed a decision tree exploration process as a form of post-classification analytic process, expanding the error-sensitivity evaluation scope from examining attribute-by-attribute individually, to cover a combined pattern of related attributes and their specific value ranges collectively in association with classification errors. Over the course of the research journey

at this stage, the proposed decision tree exploration process demonstrated its capability to measure the error-sensitivity level between specific value patterns in terms of classification rules and helped identify the most error-sensitive decision tree branches in terms of attributes and value ranges in a systematic way. It also helped examine the possible correlation between errors and such value patterns in order to gain a further understanding of the data from the value pattern and classification rule perspective and to raise stakeholders' awareness of these specific error-sensitive patterns and their implication and correlation with errors.

In the experiments to examine the proposed decision tree exploration process, a certain level of correlation between the weakest branches and the most error-sensitive attributes when testing with a number of UCI datasets was found and certain connections between specific split-point value ranges of those weakest branches and the ambiguous value ranges of those most error-sensitive attributes were identified. Other experiments designed to filter out certain attributes with a higher risk level identified from the highly error-sensitive classification rules in the form of decision tree branches also produced some supportive results in validating such an investigative method of error reduction. While the resulting improvements are not as significant as hoped, they can still be regarded as encouraging evidence in highlighting the context and correlation between error-sensitive value patterns, classification rules and errors, and potentially adding another dimension for data examination and comprehension.

In order to gain further insight into a classification model as a way to facilitate the identification and analysis of error-sensitive value patterns, a decision tree enumeration

process was developed in the next step of the research journey, as discussed in Chapter 5, to tag and map a decision tree by enumerating individual nodes from the root to all the leaves in a simple and digital way, so each node can be identified and addressed in a tree structure accordingly and uniquely, and each node's classification statistics, such as accuracy rate and error count, can also be stored and compared systematically, individually and concisely, making the identification of risky and error-sensitive nodes easier and simpler.

The results and findings from the experiments with five UCI datasets demonstrated a certain level of support for this decision tree enumeration process. For example, a long and verbiage classification rule can now be formulated and addressed as individual node IDs in a unique, numerical and contextual way for easy node-level statistical analysis, so the progress of each data sample passing each decision point along each classification path can be traced and recorded and the weakest nodes and their highly error-sensitive value patterns are now easier to identify and examine.

In addition, classification models are mostly considered as a kind of black-box. After feeding a dataset as the input of a classification process, its classification result just pops out magically as output, but the inner workflows and classification paths are hidden within the black-box. This decision tree enumeration idea may help rekindle more interest in mapping and transforming black-box classification into a more traceable and transparent process, to help understand the classification, the data and their interaction better.

This pursuit of error-sensitive value pattern identification and analysis from multiple perspectives for the goal of better understanding the data led to the development of a more dynamic and systematic construction process for a multi-dimensional data context

model built upon the transformation of a confusion matrix, as discussed in Chapter 6, to enable data value analysis and error pattern comparison based on incremental, categorical and correlational context both visually and dynamically, and specific value correlation patterns can be compared and analyzed category by category and growth stage by growth stage, side-by-side and step-by-step.

When this context construction process is applied to the UCI datasets in the experiments to test and verify its effectiveness, some meaningful and supportive results and findings can be observed. One such key observation is for samples in the true negative category, where the value variation amongst attributes appears to be more in-synch, consistent and in a narrower margin. However, for samples in the true positive, false negative and false positive category, the value variations amongst attributes appear to fluctuate erratically with much bigger margins. This clear contrast between the true negative category and the rest can be seen as a well-matched but less-poetic reflection of the Anna Karenina principle, “Happy families are all alike; every unhappy family is unhappy in its own way”.

Another interesting and meaningful observation from the experiments is the correlation between attributes and the interrelation between result categories can now be outlined and compared along the growth path of a specific lead attribute systematically and graphically. This has introduced extra dimensions into value pattern analysis and provided meaningful context when comparing value patterns and characteristics between data samples. It is all about contributing to data comparison and comprehension in an incremental, categorical and correlation way, and this is just a start.

## 7.2 Contributions

When sharing the key developments and findings from this research journey with the community for debate and discussion, it can be considered as a form of contribution aiming to generate more interest and motivation for further exploration and for the benefit of knowledge building and sharing. There are two key perspectives of this contribution that have been emphasized in this research study and the thesis, one from a philosophical perspective and one from a practical perspective.

From a philosophical perspective, no one single classification algorithm can always achieve the highest accuracy when tested on different datasets in different situations because the different processing logic within different algorithms may suit different datasets with different characteristics of data values in one specific way under a specific condition. This may explain the emergence of ensemble classification methods and discussion topics about data analysis beyond accuracy in statistical and quality management terms [WS96] [SJS06], and it is also the key reason that drives this research journey to emphasize the importance of further understanding about various aspects of a specific dataset itself in addition to the usual focus on classification theory and algorithm development. However, the soundness of this study itself in terms of its progress from ambiguity and sensitivity to transparency and contextuality may still be lacking due to insufficient theoretical proof and sophistication, so this is a contributing counter-example for further debate and discussion.

From a practical perspective, this research study introduced four independent processing models on value pattern analysis in a progressive and contextual way, to implement and complement the philosophical arguments used in the study and to demonstrate some

encouraging and supportive results and findings from the experiments with real-world data. For example, some highly ambiguous and error-sensitive attributes identified by the ambiguity and error-sensitive attribute evaluation model turn out to be also the highly significant attributes ranked by the gain ratio and information gain algorithm, and many of the multi-dimensional contexts constructed by the confusion matrix transformation model reflect the Anna Karenina principle in a consistent way.

### 7.3 Future Study

While it is easy to highlight the importance of understanding ambiguity, sensitivity and context in terms of value patterns in words, it can be rather difficult to prove this satisfactorily in the experiments because of the intangible nature of ambiguity and context. In an attempt to take this research journey of ambiguity and contextuality further, a plan for further study has been drawn up as follows:

- Evaluate more methods to identify ambiguous value range. Candidates may include the median value of the class label method and stratified weighted average method, etc.
- Evaluate alternate methods for error-sensitive attribute identification such as using the error density function
- Expand the decision tree enumeration model to map out the random forest model to keep track and illustrate the determination, selection and classification flows for transparency and better understanding and if successful, more classification models will also be considered



- Adopt other value normalization techniques and sampling techniques for context construction, such as converting the z-score of attribute values into incremental units and reconstructing deciles by stratified random sampling
- Expand the development of the context construction model in two directions, where one direction is to expand inwardly to target a specific growth stage or a particular cluster of attributes to allow for localized and detailed value pattern study and the other direction is to expand outwardly to include other contexts for a more cohesive contextual environment, such as temporal context for progression timing and territorial factors for geo-graphical and situational context

This research journey of error-sensitive value pattern analysis from ambiguity and sensitivity to transparency and contextuality is about exploring and gaining further understanding about data from various aspects, and it is hoped that this plan of further study can help initiate a new expedition for further context exploration and more knowledge discovery.

## REFERENCES

- [ABG00] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sa and L. Pereira-Leite. "SisPorto 2.0: a program for automated analysis of cardiotocograms." *Journal of Maternal-Fetal Medicine*, 9(5), pp.311-318. 2000.
- [Alp09] E. Alpaydin. "Introduction to machine learning." MIT press, 2009.
- [Bad94] M.L. Badgero. "Digitizing artificial neural networks." In *Neural Networks*, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on (Vol. 6, pp. 3986-3989). IEEE. 1994.
- [Ban60] R.B. Banerji. "An Information Processing Program for Object Recognition." *General Systems*, vol. 5, pp. 117-127. 1960.
- [BB98] E.J. Bredensteiner and K.P. Bennett. "Feature Minimization within Decision Trees." *Computational Optimization and Applications*, vol. 10, no. 2, pp. 111-126. 1998.
- [BFO84] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. "Classification and Regression Trees." Wadsworth International Group, Belmont, Calif. 1984.
- [BL13] K. Bache and M. Lichman. UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. University of California, School of Information and Computer Science, Irvine, CA. 2013.
- [Bre01] L. Breiman. "Random Forests." *Machine learning*, vol. 45, no.1, pp. 5-32. 2001.

- [Bre96] L. Breiman. "Bagging Predictors." *Machine learning*, vol. 24, no.2, pp. 123-140. 1996.
- [Cas12] G. M. Castillo. "Modelling patient length of stay in public hospitals in Mexico." (Doctoral dissertation, University of Southampton). 2012.
- [CB91] B. Cestnik and I. Bratko. "On estimating probabilities in tree pruning." *Machine Learning - EWSL-91*, vol. 482, pp. 138-150. 1991.
- [CC06] F. Cozman and I. Cohen. "Risks of semi-supervised learning." *Semi-Supervised Learning*. 2006.
- [CM98] G.A. Carpenter and N. Markuzon. "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases." *Neural Networks*, vol. 11, pp. 323-336. 1998.
- [Coc97] W.G. Cochran. "Sampling Techniques." 3rd Ed. Wiley. 1997.
- [Cov01] M.B. Covington. "Traditional Chinese medicine in the treatment of diabetes." *Diabetes spectrum*, 14(3), pp.154-159. 2001.
- [CSZ06] O. Chapelle, B. Schölkopf and A. Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]." *IEEE Transactions on Neural Networks* 20.3: 542-542. 2009.
- [DA99] A. K. Dey and G. D. Abowd. "Towards a better understanding of context and context-awareness." *International symposium on handheld and ubiquitous computing*. Springer, Berlin, Heidelberg. 1999.
- [DFL11] G. Danaei, M. Finucane, Y. Lu, G. Singh, M. Cowan, C. Paciorek, J. Lin, F. Farzadfar, Y. Khang, G. Stevens, M. Rao, M. Ali, L. Riley, C. Robinson and M.

- Ezzati. "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants.", *The Lancet*, Volume 378, Issue 9785, 2011.
- [Dia98] J. M. Diamond. "Guns, germs and steel: a short history of everybody for the last 13,000 years." Random House. 1998.
- [DLR77] A. P. Dempster, N. M. Laird and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1-38. 1977.
- [Dom99] P. Domingos. "Metacost: A general method for making classifiers cost-sensitive." *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999.
- [DX13] J. Dai and Q. Xu. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing* 13.1: 211-221. 2013.
- [Faw06] T. Fawcett. "An Introduction to ROC Analysis." *Pattern recognition letters*, 27(8), pp.861-874. 2006.
- [Fei59] E. A. Feigenbaum. "An information processing theory of verbal learning." *The RAND Corporation*, p. p.1817. 1959.
- [Fin12] E. Finn. "The advantage of ambiguity." Retrieved from <http://news.mit.edu/2012/ambiguity-in-language-0119>. January 19, 2012.

- [Fra00] E. Frank. "Pruning decision trees and lists" (Doctoral dissertation, University of Waikato). 2000.
- [FS95] Y. Freund, R.E. Schapire. "A Decision-Theoretic Generalization of Online Learning and an Application to Boosting." In Computational learning theory, pp. 23-37. 1995.
- [GE03] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection.", The Journal of Machine Learning Research, vol. 3, pp. 1157-1182. 2003.
- [Gro04] E. Grossmann. "AdaTree: Boosting a Weak Classifier into a Decision Tree." In Computer Vision and Pattern Recognition Workshop, 2004
- [GS10] I. A. Gheyas, L. S. Smith. "Feature Subset Selection in Large Dimensionality Domains." Pattern recognition, vol. 43, no.1, pp. 5-13. 2010.
- [GS88] R. P. Gorman and T. J. Sejnowski. "Analysis of hidden units in a layered network trained to classify sonar targets." Neural networks, 1(1), pp.75-89. 1988.
- [Han07] D. J. Hand. "Principles of data mining." Drug safety, 30(7), pp.621-622. 2007.
- [HFH09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. "The WEKA Data Mining Software: An Update.", SIGKDD Explorations, vol. 11, no. 1. 2009.
- [HG03] M. Hall and H. Geoffrey. "Benchmarking attribute selection techniques for discrete class data mining." IEEE Transactions on Knowledge and Data engineering 15.6: 1437-1447. 2003.

- [HGC16] A. C. Hernández, C. Gómez, J. Crespo and R. Barber. "Object detection applied to indoor environments for mobile robot navigation." *Sensors* 16, no. 8 (2016): 1180. 2016.
- [HK06] J. Han and M. Kamber. "Data mining: concepts and techniques." 2 edn, Morgan Kaufmann, San Francisco, CA. 2006.
- [HKC98] D. Heck, J. Knapp, J. Capdevielle, G. Schatz and T. Thouw. "A Monte-Carlo code to simulate extensive air showers." Report FZKA 6019. Forschungszentrum Karlsruhe. 1998.
- [HN96] P. Horton and K. Nakai. "A probabilistic classification system for predicting the cellular localization sites of proteins." In *Ismb* (Vol. 4, pp. 109-115). 1996.
- [Ho95] T.K. Ho. "Random decision forests." In *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1, pp. 278-282. 1995.
- [HT02] S. Hashemi and T. Trappenberg. "Using SVM for Classification in Datasets with Ambiguous data." *SCI*. 2002.
- [HTF01] T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference and Prediction." Springer, New York; 2001.
- [Hun62] E. B. Hunt. "Concept learning: An information processing problem." Hoboken, NJ, US: John Wiley & Sons Inc. 1962.
- [Kit78] J. Kittler. "Feature set search algorithms." *Pattern recognition and signal processing*, vol. 41, p. 60. 1978.
- [KK17] S. K. Kwak and J. H. Kim. "Statistical data preparation: management of missing values and outliers." *Korean journal of anesthesiology*, 70(4), pp.407-411. 2017.

- [KMM05] P. Klibanoff, M. Marinacci and S. Mukerji. "A smooth model of decision making under ambiguity." *Econometrica*, 73(6), pp.1849-1892. 2005.
- [Kni21] F. H. Knight. "Risk, Uncertainty and Profit." Boston: Houghton Mifflin Company. 1921.
- [Kov14] U. Kovala. "Theories of context, theorizing context." *Journal of Literary Theory*, 8(1), pp.158-177. 2014.
- [KP98] R. Kohavi and F. Provost. "On Applied Research in Machine Learning." In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Columbia University, New York, volume 30. 1998.
- [KR92] K. Kira and L.A. Rendell. "A practical approach to feature selection." *Proceedings of the ninth international workshop on Machine learning*, Morgan Kaufmann Publishers Inc., pp. 249-256. 1992.
- [Kun04] L. I. Kuncheva. "Combining Pattern Classifiers: Methods and Algorithms." John Wiley & Sons. 2004.
- [KY03] K. Kayaer and T. Yyldyrym. "Medical diagnosis on pima indian diabetes using General Regression Neural Networks." Paper presented to the International Conference on Artificial Neural Networks / International Conference on Neural Information Processing, Istanbul, Turkey. 2003.
- [Leo10] P. M. Leonardi. "Digital materiality? How artifacts without matter, matter." *First monday*, 15(6). 2010.

- [LG94] D. D. Lewis and W.A. Gale. "A sequential algorithm for training text classifiers." Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., pp. 3-12. 1994.
- [LM18] R. Ligeiro and R.V. Mendes. "Detecting and quantifying ambiguity: a neural network approach." *Soft Computing*, 22(8), pp.2695-2703. 2018.
- [LMS10] H. Liu, H. Motoda, R. Setiono and Z. Zhao. "Feature selection: An ever evolving frontier in data mining." *Feature Selection in Data Mining*, pp. 4-13. 2010.
- [Loh09] S. Lohr. "Sampling: design and analysis." Nelson Education. 2009.
- [LR57] R. Luce and H. Raiffa. "Games and decisions." Wiley, New York. 1957.
- [LWN06] Y. M. Lin, X. Wang, W. W. Ng, Q. Chang, D. S. Yeung and X. L. Wang. "Sphere classification for ambiguous data." In *Machine Learning and Cybernetics, 2006 International Conference on* (pp. 2571-2574). IEEE. 2006.
- [LY05] H. Liu and L. Yu. "Toward integrating feature selection algorithms for classification and clustering." *IEEE Transactions on knowledge and data engineering* 17.4: 491-502. 2005.
- [LYW04] C. X. Ling, Q. Yang, J. Wang and S. Zhang. "Decision trees with minimal costs.", *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, ACM New York, NY, USA. 2004.
- [MB13] W. Marx and L. Bornmann. "The emergence of plate tectonics and the Kuhnian model of paradigm shift: a bibliometric case study based on the Anna Karenina principle." *Scientometrics*, 94(2), pp.595-614. 2013.



- [MCS11] K. Monteith, J. L. Carroll, K. Seppi and T. Martinez. "Turning Bayesian Model Averaging into Bayesian Model Combination." In the 2011 International Joint Conference on Neural Networks, pp. 2657-2663. 2011.
- [Min89] J. Mingers. "An empirical comparison of pruning methods for decision tree induction" Machine learning, vol. 4, pp. 227-243. 1989.
- [MRA95] M. Mehta, J. Rissanen and R. Agrawal. "MDL-Based Decision Tree Pruning." In KDD (Vol. 21, No. 2, pp. 216-221). 1995.
- [MSW94] O. L. Mangasarian, W.N. Street and W .H. Wolberg. "Breast Cancer Diagnosis and Prognosis via Linear Programming." Mathematical Programming Technical Report. 1994.
- [Nar97] K. Narayan. "Diabetes mellitus in Native Americans: The problem and its implications." Population Research and Policy Review 16.1-2: 169-192. 1997.
- [NK91] K. Nakai and M. Kanehisa. "Expert system for predicting protein localization sites in gram - negative bacteria." Proteins: Structure, Function, and Bioinformatics, 11(2), pp.95-110. 1991.
- [NMM06] K. Nigam, A. McCallum and T. Mitchell. "Semi-supervised text classification using EM." Semi-Supervised Learning, pp. 33-56. 2006.
- [Paw98] Z. Pawlak. "Rough set theory and its applications to data analysis." Cybernetics & Systems 29.7: 661-688, 1998.
- [PBD10] K. Patel, N. Bancroft, S.M. Drucker, J. Fogarty, A.J. Ko and J. Landay. "Gestalt: integrated support for implementation and analysis in machine learning." In

Proceedings of the 23rd annual ACM symposium on User interface software and technology (pp. 37-46). ACM. 2010.

[PD00] F. Provost and P. Domingos. "Well-trained PETs: Improving probability estimation trees." 2000.

[PF01] F. Provost and T. Fawcett. "Robust classification for imprecise environments." *Machine learning*, 42(3), pp.203-231. 2001.

[PF13] F. Provost and T. Fawcett. "Data science and its relationship to big data and data-driven decision making." *Big data*, 1(1), pp.51-59. 2013.

[PFK98] F. Provost, T. Fawcett and R. Kohavi. "The Case Against Accuracy Estimation for Comparing Induction Algorithms." In *ICML* (Vol. 98, pp. 445-453). 1998.

[PHG15] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan and K. Hoagwood. "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research." *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), pp.533-544. 2015.

[Phi12] T. Phihl. "Medical Geography of the Pima Indian Reservation Diabetes Epidemic: The Role of the Gila River." *Undergraduate Honors Theses*. Paper 304, 2012.

[PW08] J. Parsons and Y. Wand. "Using cognitive principles to guide classification in information systems modeling." *MIS quarterly*, pp.839-868. 2008.

[QR89] J. R. Quinlan and R. L. Rivest. "Inferring decision trees using the minimum description length principle." *Information and computation*, 80(3), pp.227-248. 1989.

- [Qui86] J. R. Quinlan. "Induction of decision trees." *Machine learning* 1.1: 81-106. 1986.
- [Qui87] J. R. Quinlan. "Simplifying decision trees." *International Journal of Man-Machine Studies*, vol. 27, no. 3, p. 14. 1987.
- [Qui93] J. R. Quinlan. "C4. 5: programs for machine learning.", Morgan Kaufmann. 1993.
- [RDK01] M. L. Raymer, T. E. Doom, L.A. Kuhn and W.L. Punch. "Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm.", *Bioinformatics and Bioengineering Conference*, 2001, Proceedings of the IEEE 2nd International Symposium, pp. 236 - 245. 2001.
- [RL05] P. J. Rousseeuw and A. M. Leroy. "Robust regression and outlier detection." (Vol. 589). John wiley & sons. 2005.
- [RM14] L. Rokach and O. Maimon. "Data Mining with Decision Trees: Theory and Applications." World scientific. 2014.
- [Rob50] R. Robinson. "Forms and error in Plato's Theaetetus." *The Philosophical Review*, 59(1), pp.3-30. 1950.
- [Sam93] C. Sammut. "The origins of inductive logic programming: a prehistoric tale.", *Third International Workshop on Inductive Logic Programming*, ed. S. Muggleton, J. Stefan Institute, Bled, Slovenia, pp. 127-147. 1993.
- [SBR06] L. O. Schulz, P. H. Bennett, E. Ravussin, J. R. Kidd, K. K. Kidd, J. Esparza and M. E. Valencia. "Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US." *Diabetes Care*, 29(8), pp.1866-1871. 2006.

- [Sch90] R. E. Schapire. "The Strength of Weak Learnability." *Machine learning*, vol. 5, no.2, pp. 197-227. 1990.
- [SCR08] B. Settles, M. Craven and S. Ray. "Multiple-instance active learning." *Advances in neural information processing systems*. 2008.
- [SED88] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler and R. S. Johannes. "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus.", *Proc Annu Symp Comput Appl Med Care*, vol. 9, pp. 261–265. 1988.
- [See02] M. Seeger. "Learning with labeled and unlabeled data." *Institute for Adaptive and Neural Computation*. 2002.
- [Seg19] T. Segal. "Enron Scandal: The Fall of a Wall Street Darling." Retrieved from <https://www.investopedia.com/updates/enron-scandal-summary>. Last updated 2019 May-29.
- [Set09] B. Settles. "Active Learning Literature Survey." *University of Wisconsin, Madison*. 2009.
- [Sha48] C. E. Shannon. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656. 1948.
- [SIL07] Y. Saeys, I. Inza and P. Larranaga. "A review of feature selection techniques in bioinformatics.", *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517. 2007.
- [SJS06] M. Sokolova, N. Japkowicz and S. Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Springer, Berlin, Heidelberg. 2006.

- [SMW95] W. N. Street, O. L. Mangasarian and W. H. Wolberg. "An inductive learning approach to prognostic prediction." In Machine Learning Proceedings 1995 (pp. 522-530). 1995.
- [SNG13] M. K. Stein, S. Newell, R. Galliers and E. Wagner. "Classification systems, their digitization and consequences for data-driven decision making: Understanding representational quality." Proceedings of the Thirty-Fourth International Conference on Information Systems, Milan, Italy. 2013.
- [SOS92] H.S. Seung, M. Oppen and H. Sompolinsky. "Query by committee." Proceedings of the fifth annual workshop on Computational learning theory, ACM, pp. 287-294. 1992.
- [Ste99] J. Stepaniuk. "Rough set data mining of diabetes data." Foundations of Intelligent Systems: 457-465, 1999.
- [Tay96] J. R. Taylor. "An Introduction to error analysis : The Study of uncertainties in physical measurements.", 2nd edition, University Science Books, Sausalito, California. 1996.
- [TB00] T. P. Trappenberg and A. D. Back. "A classification scheme for applications with ambiguous data." In Neural Networks, 2000. Proceedings of the IEEE-INNS-ENNS International Joint Conference on (Vol. 6, pp. 296-301). IEEE. 2000.
- [TBA99] T. P. Trappenberg, A. D. Back and S. I. Amari. "A performance measure for classification with ambiguous data." BSIS Technical Report. 1999.

- [TCC13] S. Toma, L. Capocchi and G.A. Capolino. "Wound-rotor induction generator inter-turn short-circuits diagnosis using a new digital neural network." IEEE Transactions on Industrial Electronics, 60(9), pp.4043-4052. 2013.
- [TMA14] S. Tabakhi, P. Moradi and F. Akhlaghian. "An Unsupervised Feature Selection Algorithm Based on Ant Colony Optimization." Engineering Applications of Artificial Intelligence, vol. 32, pp. 112-123. 2014.
- [Tu05] Z. Tu. "Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering." In Tenth IEEE International Conference on Computer Vision, vol. 2, pp. 1589-1596. 2005.
- [VRR11] S. Visa, B. Ramsay, A.L. Ralescu and E. Van der Knaap. "Confusion Matrix-based Feature Selection." In MAICS, pp. 120-127. 2011.
- [VWN05] M. E. Valencia, E. J. Weil, R. G. Nelson, J. Esparza, L. O. Schulz, E. Ravussin and P. H. Bennett. "Impact of lifestyle on prevalence of kidney disease in Pima Indians in Mexico and the United States." Kidney International, 68, pp.S141-S144. 2005.
- [WA04] L. Wei and R.B. Altman. "An Automated System for Generating Comparative Disease Profiles and Making Diagnoses.", IEEE Transactions on Neural Networks, vol. 15. 2004.
- [WBH02] G. Williams, R. Baxter, H. He, S. Hawkins and L. Gu. "A comparative study of RNN for outlier detection in data mining." In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 709-712). IEEE. 2002.

- [WF05] I. H. Witten and E. Frank. "Data Mining: Practical machine learning tools and techniques." 2nd edn, Morgan Kaufmann, San Francisco, CA 94111. 2005.
- [WHM12] D.G. Whiting, J.V. Hansen, J.B. McDonald, C. Albrecht and W.S. Albrecht. "Machine Learning Methods for Detecting Patterns of Management Fraud." *Computational Intelligence* 28, No. 4, pp.505-527. 2012.
- [WM90] W. H. Wolberg and O. L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." *Proceedings of the National Academy of Sciences*, vol. 87, pp. 9193--9196. 1990.
- [WS96] R. Wang and D. Strong. "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems* (12:4), pp. 5–34. 1996.
- [Wu13] W. Wu. "Exploring error-sensitive attributes and branches of decision trees." *Ninth International Conference on Natural Computation (ICNC) 2013*, pp. 929-934. IEEE. 2013.
- [Wu15] 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.
- [Wu19a] W. Wu. "Weakly Supervised Learning by a Confusion Matrix of Contexts." Paper accepted for Presentation to *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2019, International Workshop on Weakly Supervised Learning: Progress and Future*. 2019.

- [Wu19b] 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.
- [WWC13] Z. Wang, J. Wang and P. Chan. "Treating type 2 diabetes mellitus with traditional Chinese and Indian medicinal herbs." *Evidence-Based Complementary and Alternative Medicine*, 2013.
- [WZ13] 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.
- [Yoo89] K. Yoon. "The propagation of errors in multiple-attribute decision analysis: A practical approach.", *Journal of the Operational Research Society*, vol. 40, no. 7, pp. 681-686. 1989.
- [YYZ10] P. Yang, Y.H. Yang, B. Zhou and A. Zomaya. "A Review of Ensemble Methods in Bio-informatics." *Current Bioinformatics*, vol. 5, no.4, pp. 296-308. 2010.
- [Zhu05] X. Zhu. "Semi-supervised learning literature survey." University of Wisconsin, Madison. 2005.
- [ZWY08] J. Zhu, H. Wang, T. Yao and B. Tsou. "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification." *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1. 2008.