ACKNOWLEDGEMENTS

It is my pleasure to acknowledge the help and encouragement of all who have supported and assisted me during this thesis research. Without their guidance, advice and motivation, I would have never been able to accomplish the work of this thesis.

Foremost, I would like to thank the Vice Chancellor, Registrar and Dean Research and Consultancy Division, BITS Pilani for giving me an opportunity to do this research and enhance my professional career.

I would like to express my deepest gratitude to the members of the Doctoral Advisory Committee BITS Pilani, for their valuable comments on my research proposal which have paved the correct path for pursuing my thesis.

I would like to thank **Prof. Dr. R.N. Saha,** Director, BITS Pilani, Dubai Campus for his encouragement and facilities he has provided to carry out this research work. I would like to thank **Senior Prof. Dr. R.K. Mittal**, Director, Special Projects, BITS Pilani and Former Director BITS Pilani, Dubai Campus for the constant encouragement, facilities and support he has provided to carry out this research work.

I am grateful to Dr. M. Ramachandran, Dr. G. Vijaya, Dean Research and Consultancy Division, BITS Pilani, Dubai Campus, Dr. K. K. Singh, the Doctoral Advisory Committee, BITS Pilani, Dubai Campus, Dr. M. Madiajagan, the Doctoral Research Committee Convenor for their constant monitoring and advisory role throughout my research work. My sincere gratitude to Dr. Sameen S. Fathima, who introduced me to this research area.

I would like to express my deepest gratitude to my supervisor **Dr. V. Santhosh Kumar,** Assistant Professor, BITS Pilani, Dubai Campus for his research guidance, advice, technical inputs and constant encouragement during the course of this research. Our valuable discussions encouraged me to complete this work. I would like to thank Dr. S. Vadivel, HOD CS and Dr. B. Vijaya Kumar, Associate Professor, CS, BITS Pilani, Dubai Campus and all of my department colleagues for their valuable queries during my presentations. I would also like to thank Dr. Jagadish Nayak, Assistant Professor, BITS Pilani, Dubai Campus for his valuable inputs and guidance to complete my research.

I would also thank the IT Support Team and Dr. K.N. Sheshadri, Librarian, BITS Pilani, Dubai Campus for extending excellent technical facilities and providing library resources on time that helped me to complete this thesis. I acknowledge all my colleagues, department technical staff for their support to complete my research.

I am very much thankful to my family for providing me continuous moral support and encouragement. Above all, I sincerely thank the Almighty for giving me an opportunity, strength and support to carry out my research work to the best of my abilities.

J. ALAMELUMANGAI

2008PHXF401U

ABSTRACT

Recent studies have shown that the significance of data-driven decision making and the notion of "Big-Data" are recognized broadly. Advances in data collection and storage technologies have made large organizations to accumulate vast volumes of data with ease in digital format. But extracting useful information from such massive volumes of data using traditional data analysis techniques is quite challenging. The various factors that impede the progress of "Big-Data" analysis in all phases during its analysis are heterogeneity, scale, complexity, timeliness and privacy problems. Also, it is important to decide which data to retain and which to discard during the data acquisition phase itself. Data Mining is a solution to overcome much of the challenges faced by the traditional data analysis techniques in such situations. Data Mining is a technology that combines traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It is a process of automatically discovering novel and useful patterns that are hidden in large data repositories.

Traditional data mining techniques have been developed mainly for structured data types. But much of the data available today is not in the native structured format; for example, tweets and blogs have large pieces of unstructured text data; Images and videos are structured for storage and display, but not for semantic content and search; Web pages have data of different formats embedded in them and hence are semi-structured. Transforming such contents into a structured format for later analysis is itself a major challenge.

The objective of this thesis is to design and implement algorithms for improving subject based classification of web page and medical image data sets using data mining techniques. Classifying a web page/a medical image into one of a pre-defined category is known as web page/medical image classification. With millions of new web pages being added each day the volume of WWW will defeat any conceivable team of human classifiers. Hence automatic classification of web pages is required. Such automated tools also help the search engines to construct web directories and hence make a relevant and quick retrieval of information for the user query. The ever increasing amounts of patient data in the form of medical images, imposes new challenges to clinical routine, such as diagnosis, treatment and monitoring. Medical data mining refers to the process of transforming raw imaging data using knowledge-based data mining algorithms into clinically relevant information. The target mining model enables a physician to spend less time in spending on the image volumes to extract the clinical information in it, while improving the diagnostic accuracy.

The data sets are classified using the content embedded in them and hence the present classification framework is subject-based. Algorithms for the various pre-processing steps namely feature extraction, feature selection and discretization are designed and implemented to improve the predictive accuracy of the classification model. Two new feature selection methods are designed and implemented in this thesis. First is a hybrid model using CFS (Correlation Based Feature Selection) and Decision Tree. Second is a novel method using Ward's minimum variance measure to identify clusters of redundant features. This research also presents two new classification models for web page and image data sets. The first model is a probabilistic method (Probabilistic Web Page Classifier and Probabilistic Medical Image Classifier). Experimental analysis is done for both binary-class and multi-class classification.

The results of the research identify that the present feature selection method helps in improving the predictive accuracy and modeling the classifier with less number of features. The performance of many of the binary-class classifiers is better in the discrete domain than in continuous domain. The predictive accuracy of the present MKNN classification framework with both web page and image data sets is better than the traditional KNN for both binary-class and multi-class classification. The performance of the present PWPC and PMIC is also better than many of the existing classifiers for both binary-class and multi-class classification. Its performance depends on the discretization algorithm that is run prior to classification.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS……………………………………………………………………..i

ABSTRACT……………………………………………………………………………..iii

LIST OF TABLES……………………………………………………………….....…xi

LIST OF FIGURES ……………………………………………………………….....xv

LIST OF ABBREVIATIONS ………………………………………………………......xvii

CHAPTER 1 INTRODUCTION……………………………………………………….1

1.1 Challenges of Traditional Data Analysis Methods……………………..……...............1

1.2 Objective, Scope and Limitations ……………..…………………….….……............3

1.2.1 Web Page Classification……………………………………………………..4

1.2.1.1 Motivation ……………………………………………………….…4

1.2.1.2 Architecture of a Web Page Classification Model ………….…………6

1.2.1.3 Web Page Representation ……………...………………………………8

1.2.2 Medical Image Classification…………………………………………………12

1.2.2.1 Motivation ………………………………………………………….12

1.2.2.2 Architecture of a Medical Image Classification Model ………………....12

1.2.2.3 Image Representation ……………………………………………...14

1.3 Background Work ………………………………………………………………….15

1.3.1 Knowledge Discovery and Data Mining ……………….……………………..15

1.3.2 Data Pre-Processing ……………………………….…………………………17

1.3.3 Data Mining Tasks …………………………………………….…………...24

1.3.4 Applications of Data Mining ……………………………………………………25

1.3.5 Data Mining Tools …………………..………………………………………...29

1.3.6 Supervised Machine Learning ……………………………………………………..30

1.3.7 Naïve Bayes Classification ……………………………………………………...33

1.3.8 K Nearest Neighbor Classification ……………………………………………...34

1.3.9. Association Rule Mining ……………………………...…………………...36

1.4 Research Gap …………………………………………………...……………….37

1.5 Organization of the Thesis ………………………………………………………39

CHAPTER 2 RELATED WORK …………...…………………………………………..41

2.1 Survey of Web Page Classification Methods ………….………………………...41

2.2 Survey of Medical Image Classification Methods ……….……………………...51

2.3 Summary ………………………………….……………………………………..53

CHAPTER 3 ALGORITHMS FOR IMPROVING SUBJECT BASED CLASSIFICATION OF WEB PAGE AND MEDICAL IMAGE DATA ….……………………54

3.1 Present Framework for Web Page Classification WPC ……………………………..54

3.1.1 Feature Extraction and Web Page Representation ………………………………54

3.1.2 Feature Selection ……………………………………………………………..58

3.1.2.1 Feature Selection using CFS and C4.5 ………………….....................59

3.1.2.2 Feature Selection Framework using Ward's Minimum Variance Measure ………………………………………………………………62

3.1.3 Data Tuning ……………………………………………………………………..67

3.1.4 Discretization …………………………...……………………………………68

3.1.5 The Web Page Classification Algorithms……………………………………..71

3.1.5.1 The Probabilistic Web Page Classifier PWPC ………………………..71

3.1.5.2 The Modified k Nearest Neighbor Classification MKNN ……………74

3.2 Present Framework for Medical Image Classification MIC ...…………………………..….78

3.2.1 Image Preprocessing ……………………………………………………78

3.2.2 Feature Extraction ……………………………………….……………79

3.2.3 Feature Selection …………….………………………………………..80

3.2.4 Feature Discretization ……………………………………………….…...80

3.2.5 Classification Algorithms for Medical Images ……………………….........81

3.3 Performance Evaluation Metrics and Methods ………………………………………82

3.4 Summary ……………………………………………...……………………………85

CHAPTER 4 RESULTS AND DISCUSSION……………...…………………………………87

4.1 Experimental Setup ………………………………………………………….….... 87

4.1.1 Data Set Description …………………………..…………………………88

4.1.2 Implementation Overview …………………………………………………89

4.2 Web Page Classification …….……………………..…………………………....90

4.2.1 Binary Class Web Page Classification ………………………………..………91

4.2.1.1 Feature Extraction ……………………...……………………………..91

4.2.1.2 Feature Selection …………………...……………………………………94

4.2.1.3 Feature Discretization …….…………………………………………...105

4.2.1.4 The Probabilistic Web Page Classifier PWPC ………..………………. 111

4.2.1.5 The Modified k Nearest Neighbor MKNN Classifier ….………………114

4.2.2 Multi Class Web Page Classification ……………………………………………….117

4.3 Medical Image Classification ………………………………………………………….119

4.3.1 Binary Class Medical Image Classification ……………………………………119

4.3.1.1 Image Feature Extraction and Feature Selection ………..…………….119

4.3.1.2 Image Feature Discretization ……………………………………..……124

4.3.1.3 The Probabilistic Medical Image Classifier PMIC and MKNN

Medical Image Classifier ………………………………………..126

4.3.2 Multi Class Medical Image Classification …………………………………….127

4.4 Area under the Curve AUC for MKNN Medical Image Classification …………..…….132

4.5 Summary ……………………………….………………………………………136

CHAPTER 5 CONCLUSION ……………………………..……………………………..138

5.1 Specific Contributions ……………………………………………………………139

5.2 Future Scope of Work ………………………………..……………………………141

REFERENCES ……………………………………………………………………….143

APPENDICES

APPENDIX A: Sample Web Pages of Each Category……………………………..……....161

APPENDIX B: Brief Description of the Implementation of the Present Algorithms…………165

APPENDIX C: System Configuration………………………………………………………173

APPENDIX D: The 70-30 File after Feature Selection by CFS in Sparse ARFF for Binary

Class WPC …………………………………………………………………174

APPENDIX E: The 70-30 File after Feature Selection by the Hybrid Model in Sparse ARFF

for Binary Class WPC ……………………………………………...………….181

APPENDIX F: Features Selected by the Hybrid Model from each Input File for Binary Class

WPC …………………………………………………………………………..186

APPENDIX G: Features Selected by the Ward’s Method from each Input File for Binary Class WPC………………………………..……………………………….187

APPENDIX H: The Sparse ARFF after Feature Selection Using CFS for Multi Class WPC ..188

APPENDIX I: The 30-30 File after Feature Selection by CFS in ARFF for Binary Class MIC..………………………………………………………………...210

APPENDIX J: The 30-50-30 File after Feature Selection by CFS in ARFF for Multi Class MIC..221

LIST OF PUBLICATIONS AND PRESENTATIONS………………..…………………244

BRIEF BIOGRAPHY OF THE CANDIDATE ………………….………………………246

BRIEF BIOGRAPHY OF THE SUPERVISOR……………….…………………………247

LIST OF TABLES

Table 1.1 Instances with Known Class Labels …………………….………………………....31

Table 3.1 The Confusion Matrix ……………….…………………………………………..83

Table 4.1 Results after Feature Extraction for Binary Class WPC ..…..………………………92

Table 4.2 Classification Accuracy after Feature Extraction for Binary Class WPC …………...93

Table 4.3 Number of Selected Features by each Feature Selection Algorithm for Binary Class WPC……………………………………………………………..………........96

Table 4.4 % Reduction in the Number of Features by the Present Hybrid Model for Binary Class WPC……………………………………..……………...……......97

Table 4.5 % Classification Accuracy of Binary Class WPC with Numeric Features after Feature Selection and Data Tuning ……………………………………………………….98

Table 4.6 Wards Cluster of Redundant Features Formed and the Final Selected Features …….99

Table 4.7 Comparison of the Wards Feature Selection with Other Feature Selection Methods for Binary Class WPC …………………………………………….………………100

Table 4.8 Classification Accuracy of KNN Classifier on the Features Selected by each Feature Selection Method for Binary Class WPC …………….………………………….102

Table 4.9 Classification Accuracy of NN Classifier on the Features Selected by each Feature Selection Method for Binary Class WPC …………………..……………………102

Table 4.10 Classification Accuracy of Boosting Classifier on the Features Selected by each Feature Selection Method for Binary Class WPC …………………...……………103

Table 4.11 Classification Accuracy of SVM Classifier on the Features Selected by each Feature Selection Method for Binary Class WPC …….…………………………...104

Table 4.12 Classification Accuracy of NB Classifier on the Features Selected by each

Feature Selection Method for Binary Class WPC …………..…………….…………..104

Table 4.13 Classification Accuracy of J48 classifier on the Features Selected by each Feature Selection Method for Binary Class WPC …….………...……………………105

Table 4.14 Web Page Data Set before and after Discretization ……….…………………..106

Table 4.15 The Number of Intervals Identified for each Feature by the Present Discretization Method for Binary Class WPC ……………………………………………………106

Table 4.16 Classification Accuracy of Binary Class WPC with Features Discretized by the Present Method ……………………………………..……...……………….107

Table 4.17 Classification Accuracy by Simple Binning for Binary Class WPC ………………109

Table 4.18 Modeling Time of Binary Class WPC in msecs with Numeric Features.. ….…….110

Table 4.19 Modeling time of binary class WPC in msecs with Features Discretized by the Present Method ……………………………..…………………………..…..110

Table 4.20 Performance Comparison of Classifiers in Numeric and Present Discrete Domain for Binary Class WPC …………………..…………….………………………111

Table 4.21 Comparison of Classification Accuracy with the Present PWPC Classifier for Binary Class WPC Using Simple Binning ……………………………………112

Table 4.22 Accuracy of the PWPC Classifier on the Present Discretized Features and Simple Binning Discretized Features for Binary Class WPC …………….…………..…..114

Table 4.23 Comparison of Classification Accuracy between KNN and MKNN for Binary Class WPC …………………………………………………………….…………...115

Table 4.24 Comparison of Classification Accuracy of the MKNN Binary class Web Page Classifier with other Existing Classifiers …………………………….………116

Table 4.25 Results after Feature Extraction and Selection by CFS for Multi Class WPC ……117

Table 4.26 Comparison of Classification Accuracy with various Features for Multi Class WPC…………………………………………………………………………….. 118

Table 4.27 Performance Comparison of the PWPC and MKNN for Multi Class WPC with other Classifiers ……………………………………………………...…………..… 118

Table 4.28 Results after Feature Extraction and Feature Selection for Binary Class MIC…...120

Table 4.29 Classification Accuracy after Image Feature Extraction for Binary Class MIC …,120

Table 4.30 Classification Accuracy after Feature Selection by CFS ……………..................121

Table 4.31 Comparative Analysis of Classification Accuracy for Binary Class MIC with various Features ……………………………………………………………………124

Table 4.32 Comparative Analysis of Binary Class MIC with various Features Using AUC 125

Table 4.33 Comparative Analysis of Classification Accuracy of Binary Class MIC with KNN and MKNN …………………………………...…………………………...…. 126

Table 4.34 Comparative Analysis of Classification Accuracy of PMIC with other Classifiers for Binary Class MIC …………………..……………………..………………… 127

Table 4.35 Results after Feature Extraction and Feature Selection for Multi Class MIC ...….128

Table 4.36 Classification Accuracy after Feature Extraction for Multi Class MIC …………..128

Table 4.37 Classification Accuracy after Feature Selection for Multi Class MIC ……..……. 128

Table 4.38 Comparative Analysis of Classification Accuracy for Multi Class MIC with various Features …………………………………………………...………………..129

Table 4.39 Comparative Analysis of AUC for Multi Class MIC with various Features……...130

Table 4.40 Comparative Analysis of Classification Accuracy of KNN and MKNN for Multi Class MIC ……………..………………………………………………….……131

Table 4.41 Comparative Analysis of Classification Accuracy of PMIC with other Classifiers for

Multi Class MIC …………………………………………………………………......132

Table 4.42 TPR and FPR for Binary Class MIC …………………..………………….………133

Table 4.43 Confusion Matrix for Binary Class MIC ……………………………......………....133

Table 4.44 TPR and FPR for Multi Class MIC ………...……………………………..…….134

Table 4.45 Confusion Matrix for Multi Class MIC ………………………………….………...134

LIST OF FIGURES

Fig 1.1 Searching the WWW……………………………………..……………………………4

Fig 1.2 Web Page Classification …………………………………….………………………6

Fig 1.3 Architecture of a Web Page Classification Model ………………..……..……………...7

Fig 1.4 A Web Page Represented using Bag-of-words in the Vector Space Model ……………10

Fig 1.5 Web Page Feature Matrix ………………………………………..………………..10

Fig 1.6 Architecture of a Medical Image Classification Model ………………..…….………13

Fig 1.7 Image Feature Matrix ……………………………………………….………………..14

Fig 1.8 Knowledge Discovery in Databases………………. … ……………..…………………16

Fig 1.9 Various Forms of Data Preprocessing ……………..……………………………………18

Fig 1.10 The Feature Subset Selection Process ………………………………………………21

Fig 1.11 Pure and Impure Bins ………………………………………….…………..........22

Fig 1.12 Data Mining Tasks ………………………….……………………………………25

Fig 1.13 The Weather Data Set in ARFF ……………………….…………………………….29

Fig 3.1 The Present Framework for Web Page Classification …………..……….……………..57

Fig 3.2. The Present Framework for Medical Image Classification …………….……………...79

Fig 3.3. The Training Phase of the Present Framework for MIC ………..……………...……..81

Fig 3.4. The Testing Phase of the Present Framework for MIC …………………...……………81

Fig 3.5 The ROC Graph …………………………………………………………………………85

Fig 4.1 Sample Medical Images of Normal Category …………………………...………….. 89

Fig 4.2 Sample Medical Images of Moderate Category ………………...……………………90

Fig 4.3 Sample Medical Images of Severe Category ……………………………………………90

Fig 4.4 Feature Selection using J48 ……………………………..…………………………94

Fig 4.5 Number of Features Selected by the Present Hybrid Method for Binary Class WPC ……………………………………………………………………………..…96

Fig 4.6 Comparison of the Wards Feature Selection with other Methods for Binary Class WPC ………………………………………………………………………….…………101

Fig 4.7 Comparison of the Modeling Time for NN Classifier with various Features for Binary Class WPC ……………………………………………………...……………..103

Fig 4.8 Comparison of % Classification Accuracy of Binary Class WPC in Numeric and Present Discrete Domain …………..….…………………………...………..….108

Fig 4.9 Comparison of % Classification Accuracy of Binary Class WPC with various Features …………………………………………………………...………...109

Fig 4.10 Comparison of the Performance of the PWPC using Simple Binning with other Classifiers for Binary Class WPC ……………………………….………..……..113

Fig 4.11 Comparison of Average Classification Accuracy between KNN and MKNN for Binary Class WPC ……………………………………….…………...……….115

Fig 4.12 Image File after Feature Extraction in ARFF for Binary Class MIC …….……….122

Fig 4.13 The Image File after Feature Selection in ARFF for Binary Class MIC …………....123

Fig 4.14 The Image File after Discretization for Binary Class MIC ……………...………...123

Fig 4.15 ROC Graph for Binary Class MIC …………………………..………….....……...134

Fig 4.16 ROC Graph for Multi Class MIC ……………………………..…………………...136

LIST OF ABBREVIATIONS

WWW	–	World Wide Web
WPC	–	Web Page Classification
HTML	–	Hyper Text Markup Language
TF	–	Term Frequency
DF	–	Document Frequency
IDF	–	Inverse Document Frequency
MIC	–	Medical Image Classification
KDD	–	Knowledge Discovery in Databases
MDL	–	Minimum Description Length
ARFF	–	Attribute Relation File Format
NB	–	Naïve Bayes
ML	–	Machine Learning
KNN	–	K Nearest Neighbor
DT	–	Decision Tree
CFS	–	Correlation Based Feature Selection
URL	–	Uniform Resource Locator
PCA	–	Principal Component Analysis
UCI	–	University of Irvine
SVM	–	Support Vector Machine
EM	–	Expectation Maximization
CART	–	Classification and Regression Trees
GA	–	Genetic Algorithms

BoW	-	Bag of Words
AI	–	Artificial Intelligence
PSO	–	Particle Swarm Optimization
FP-tree	–	Frequent Pattern tree
MKNN	–	Modified K Nearest Neighbor
ROC	–	Receiver Operating Characteristics Curve
AUC	–	Area under the Curve
TP	–	True Positive
TPR	–	True Positive Rate
FP	-	False Positive
FPR	–	False Positive Rate
MLP	–	Multilayer Perceptron
RBF	–	Radial Basis Function