A DIRECT ENSEMBLE CLASSIFIER FOR LEARNING IMBALANCED MULTICLASS DATA

SAMRY @ MOHD SHAMRIE SAININ

THESIS SUBMITTED IN FULLFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF ENGINEERING AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYSIA SABAH

2013

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL : A DIRECT ENSEMBLE CLASSIFIER FOR LEARNING IMBALANCED MULTICLASS DATA

IJAZAH: DOKTOR FALSAFAH (SAINS KOMPUTER)

Saya SAMRY @ MOHD SHAMRIE SAININ, Sesi pengajian 2009-2013, mengaku membenarkan tesis ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah membenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan untuk membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/)

[ ] SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

[ ] TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

[/] TIDAK TERHAD

Disahkan oleh,

(Tandatangan Penulis)

(Tandatangan Pustakawan)

(DR. RAYNER ALFRED)
Penyelia

Tarikh: 20 Disember 2013

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries and references, which have been dully acknowledged.

6 September 2013

Samry @ Mohd Shamrie Sainin
PK20088357

CERTIFICATION

NAME : **SAMRY @ MOHD SHAMRIE SAININ**

MATRIC NO. : **PK20088735**

TITLE : **A DIRECT ENSEMBLE CLASSIFIER FOR LEARNING IMBALANCED MULTICLASS DATA**

DEGREE : **DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)**

VIVA DATE : **17 MAY 2013**

DECLARED BY

1. **SUPERVISOR**
   Dr. Rayner Alfred

Signature

ACKNOWLEDGEMENT

First and foremost, in the name of Allah the Most Gracious and the Most Merciful. I would like to extend my deepest praise to Allah S.W.T for the blessings and guidance in completing this research and thesis.

I would like to express my deepest appreciation to my supervisor, Dr. Rayner Alfred for all his advices, guidance and support in this interesting research work that lead to the completion of this thesis. You are the best supervisor.

The Ministry of Higher Learning and Universiti Utara Malaysia for the financial support throughout my studies, and also Universiti Malaysia Sabah for the resources and facilities provided.

My wife, Suraya Alias, my sons, Aiman Haziq, new born Aish Hafiy and daughter Damia Hana, for the love, trust, inspiration and great understanding that had given me.

My parents Sainin Ghane and Rosnah Moringking, and my family for being supportive towards my studies.

And lastly, all of my friends whether in the academic circle or outside of my studies for their understanding, advices and support. This research will not be completed without you.

Samry @ Mohd Shamrie Sainin
6 September 2013

ABSTRACT

A traditional direct single classifier can be easily applied to solve a multiclass classification problem. However, the performance of a single classifier is decreased with the existence of imbalanced data in multiclass classification tasks. Thus, an ensemble of classifiers is one of the methods used to solve multiclass classification tasks. In this thesis, the problem of learning from imbalanced multiclass data classification is studied. In the multiclass classification problem, decision can be estimated not only by the final single class label, but also by other appropriate class. Many real-world multiclass classification problems can be represented into a setting where non-crisp label need to be observed. An in-depth review and method to solve this special learning task is explained in this thesis. An alternative ensemble learning framework called Direct Ensemble Classifier for Imbalance Learning (DECIML) is proposed combining the advantages of existing single classifiers and ensemble methods and strategies. The learning framework consists of ensemble learning and decision combiner model with general supervised learning algorithms as base learner. Feature selection is also applied in DECIML in order to increase the performance of the ensemble learning. In order to facilitate the experiments and future research on the imbalanced multiclass problem, a standard pool of benchmark data is created, which consists of 16 datasets with different degrees of imbalanced ratio and 4 datasets for imbalanced multiclass with feature selection purposes. The benchmark data is used to evaluate and compare the proposed frameworks with several ensemble methods, such as bagging and adaboost. The DECIML with feature selection is also evaluated and compared with methods named CFsSubsetEval and Filteredsubseteval. The results obtained show that the proposed learning frameworks are comparable to other methods. In addition, the selected benchmark data, experiments and the results are useful for future research on the imbalanced multiclass classification problem. Furthermore, the DECIML framework was applied to the real world leaf classification problem based on the shape features. Extensive experiments and results show that the DECIML method does provide a promising performance in imbalanced multiclass with highly noisy data.

ABSTRAK

A DIRECT ENSEMBLE CLASSIFIER FOR LEARNING IMBALANCED MULTICLASS DATA

Algoritma pengelasan tunggal tradisional boleh digunakan dengan mudah secara langsung untuk pelbagai masalah klasifikasi berbilang-kelas. Walau bagaimanapun, prestasi pengelas tunggal akan menurun dengan kewujudan ketidakseimbangan dalam tugas klasifikasi berbilang kelas. Oleh itu, kombinasi pengelas adalah salah satu kaedah dalam tugas klasifikasi berbilang-kelas untuk masalah ketidakseimbangan dalam perlombongan data dan pembelajaran mesin. Dalam tesis ini, masalah pembelajaran dari klasifikasi data berbilang-kelas tidak seimbang dikaji. Dalam masalah pengelasan berbilang-kelas, keputusan boleh dianggarkan bukan sahaja oleh label kelas akhir tunggal, tetapi kelas yang sesuai yang lain. Klasifikasi masalah berbilang-kelas dalam dunia sebenar kebanyakannya boleh diwakilkan menggunakan label bukan tunggal yang perlu dipatuhi. Suatu kajian semula yang mendalam dan kaedah untuk menyelesaikan tugas pembelajaran khas dijelaskan dalam disertasi ini. Rangka kerja alternatif bagi kombinasi pembelajaran yang dikenali sebagai Kombinasi Pengelas Pembelajaran Ketidakseimbangan Berbilang Kelas Secara Langsung (DECIML) dicadangkan berdasarkan kepada kelebihan pengelas tunggal yang sedia ada dan kaedah serta strategi kombinasi. Rangka kerja pembelajaran ini terdiri daripada kombinasi pembelajaran dan penggabung keputusan model dengan algoritma pembelajaran terselia sebagai pembelajar asas. Satu lagi rangka kerja pembelajaran ialah menggabungkan DECIML dan pemilihan ciri untuk meningkatkan prestasi kombinasi pembelajaran. Bagi memudahkan ujikaji dan kajian akan datang untuk data berbilang-kelas tidak seimbang, satu senarai piawai data sebagai tanda aras diwujudkan, yang mana terdiri daripada 16 set data dengan darjah nisbah ketidakseimbangan yang berbeza dan 4 dataset untuk berbilang-kelas tidak seimbang untuk tujuan pemilihan ciri. Data penanda aras ini digunakan untuk menilai dan membandingkan rangka kerja yang dicadangkan dengan beberapa kaedah kombinasi, seperti bagging dan adaboost. DECIML dengan pemilihan ciri juga dinilai dan dibandingkan dengan kaedah seperti CFsSubsetEval dan Filteredsubseteval. Hasil kajian menunjukkan bahawa rangka kerja pembelajaran yang dicadangkan adalah setanding dengan kaedah lain. Di samping itu, data tanda aras yang dipilih, eksperimen dan keputusan boleh digunakan untuk penyelidikan masa depan dalam masalah klasifikasi berbilang-kelas tidak seimbang. Di samping itu, rangka kerja DECIML telah digunakan untuk klasifikasi dunia sebenar, masalah klasifikasi daun berdasarkan ciri-ciri bentuk. Ujikaji yang mendalam dan keputusan yang diperolehi menunjukkan bahawa kaedah DECIML memberikan prestasi yang baik dalam masalah berbilang-kelas tidak seimbang dengan data yang sangat bising. Oleh itu, penemuan menarik daripada keputusan eksperimen adalah sumbangan kajian mengenai masalah pembelajaran ini.

TABLE OF CONTENTS

	Page
TITLE	i
DECLARATION	ii
CERTIFICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
ABSTRAK	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
ABBREVATION	xvi
SYMBOL	xxi
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	5
1.3 Research Questions	8
1.4 Objectives of Study	8
1.5 Significance of Study	9
1.6 Scope of Study	9
1.7 Organization of the Thesis	9
CHAPTER 2: LITERATURE REVIEW	11
2.1 Introduction	11
2.2 General Multiclass Classification Problem and Methods	11
2.3 Imbalanced Multiclass Classification Problem	15
2.4 Evaluation Metric for Imbalanced Classification	17
2.5 General Imbalanced Classification Methods	21
2.5.1 Data-level Methods	22
2.5.2 Algorithms-level Approach	24
2.6 Ensembles of Classifiers	27

2.6.1 Brief History 27
2.6.2 Why are Ensembles better than Single Classifier? 29
2.6.3 Ensemble Construction 30
2.6.4 Ensemble Combination Methods 32
2.6.5 Ensemble based Naïve Bayes and k-Nearest Neighbor 35
a. Naïve Bayes 36
b. K-nearest neighbors 38
2.6.6 Related work in Naïve Bayes and k-Nearest Neighbor for Imbalance Problem 41
2.6.7 Related works of individual ensembles using naïve Bayes and k-nearest neighbors 43
2.6.8 Related works of ensembles combining naïve Bayes and k-nearest neighbors 44
2.7 Summary 46

CHAPTER 3: THE DECIML FRAMEWORK 47
3.1 Introduction 47
3.2 The DECIML Framework 47
3.3 DECIML Implementation Example 56
3.3.1 Training Phase 56
3.3.2 Testing Phase 57
3.4 Advantages of DECIML 61
3.5 Summary 63

CHAPTER 4: EXPERIMENTAL STUDIES ON DECIML 64
4.1 Introduction 64
4.2 Base-learner and Algorithm Selection 64
4.3 Dataset Selection 65
4.4 Experimental Setup 69
4.5 Experiment Results 72
4.5.1 Direct Single Classifier 72
4.5.2 AdaBoostM1, Bagging and Random Forest 74
4.5.3 DECIML Ensembles (NB+1NN and NB+kNN) 78
4.5.4 DECIML Influencing Factors 86
4.5.5 DECIML Ensembles and Sampling with Replacement 90
4.6 Summary 96

CHAPTER 5: EXPERIMENTAL STUDIES ON DECIML WITH FEATURE SELECTION 98

5.1 Introduction 98

5.2 Feature Selection in Imbalanced Multiclass Classification 98

5.2.1 Ensemble of Naïve Bayes and k-Nearest Neighbor 98

5.2.2 Genetic-Based Feature Selection Algorithms 100

5.3 Feature Selection for DECIML 105

5.3.1 Filter based Feature Selection 105

5.3.2 Wrapper based Feature Selection using Genetic Algorithm and DECIML 108

5.4 Experimental Setup 113

5.5 Experimental Results 116

5.5.1 Overall Performance of Filter based DECIML Feature Selection 116

5.5.2 Overall Performance of Wrapper based DECIML Feature Selection 119

5.5.3 Comparison With Other Feature Selection Methods 120

5.6 Summary 123

CHAPTER 6: DECIML APPLICATION ON LEAF DATA 125

6.1 Introduction 125

6.2 Background 125

6.3 Image Processing for Initial Feature Extraction 127

6.3.1 Basic Image Processing 127

6.3.2 Half Image Acquisition 131

6.4 Feature Extraction 131

6.4.1 Shape Token 132

6.4.2 Angle Features 132

6.5 Experimental Design 135

6.6 Experiment Results 136

6.7 Summary 142

CHAPTER 7: CONCLUSION AND FUTURE WORK 143

REFERENCES 153

GLOSSARY 173

LIST OF PUBLICATIONS 178

LIST OF FIGURES

		Page
Figure 2.1	Block diagram of the feature selection process	24
Figure 2.2	General classifier ensembles construction	32
Figure 3.1	General flow in DECIML framework	49
Figure 3.2	Illustrative effect of Loss Function using k=3	52
Figure 3.3	OR-tree rules for predicting $C(x)$	53
Figure 4.1	Experimental Framework on Weka	70
Figure 4.2	Experimental Framework for DECIML	70
Figure 4.3	F-Measure comparison of average performance on 16 benchmark dataset	84
Figure 4.4	G-means comparison of average performance on 16 benchmark dataset	84
Figure 4.5	MCC comparison of average performance on 16 benchmark dataset	85
Figure 4.6	Number of classes in data (sorted) and the F-measure	87
Figure 4.7	Number of feature in data (sorted) and the F-measure	87
Figure 4.8	Number of examples in data (sorted) and the F-measure	88
Figure 4.9	Imbalance ratio in data (sorted) and the F-measure	88
Figure 5.1	Flow diagram of the wrapper DECIMLFS	110
Figure 6.1	Leaf shapes as reference to symmetrical characteristic	127
Figure 6.2	Flow diagram of the half leaf scheme	128
Figure 6.3	Prewitt edge detection algorithm	129
Figure 6.4	Thinning algorithm	130
Figure 6.5	A is the full leaf and B is a half leaf edge detection with boundary and tokens (angle) for feature extraction	131

Figure 6.6	Different distance effects on the number of tokens assigned to the boundary of the leaf image. (A) More tokens if distance is 1 (configured as 10) and (B) less tokens if distance is 3 (configured as 30)	132
Figure 6.7	tion of the processed leaf image and representation of feature extraction, θ, hypotenuse angle for token P1 and P2	133

LIST OF TABLES

		Page
Table 2.1	Confusion matrix for a general binary class classification task	17
Table 3.1	DECIML algorithm	48
Table 3.2	Class Center NNDM algorithm	51
Table 3.3	OR-tree prediction combiner	54
Table 3.4	Data from Autos (provided by KEEL) with selected features using Weka feature selection filter	55
Table 3.5	Naïve Bayes prior probabilities	56
Table 3.6	CNNDM after some loss function (with loss function $k = 5$) is applied and represented as CNNDM[i][j]	57
Table 3.7	Naïve Bayes in DECIML prediction and classification	57
Table 3.8	k-NN step for DECIML prediction and classification	58
Table 3.9	Steps in OR-tree prediction combiner	60
Table 3.10	DECIML prediction and classification	61
Table 4.1	Benchmark data reference, domain, and result from other work	68
Table 4.2	Benchmark Data Characteristics	69
Table 4.3	Benchmark data (Training and Testing)	71
Table 4.4	Classification performance (F-Measure) (based on Weka)	73
Table 4.5	Classification performance (G-Means) (based on Weka)	73
Table 4.6	Classification performance (MCC) (based on Weka)	74
Table 4.7	Classification Performance (F-Measure) of AdaBoostM1 and Boosting on 16 Benchmark Datasets (based on Weka)	75
Table 4.8	Classification Performance (G-Mean) of AdaBoostM1 and Bagging on 16 Benchmark Datasets (based on Weka)	76

Table 4.9 Classification Performance (MCC) of AdaBoostM1 and Bagging on 16 Benchmark Datasets (based on Weka) 77

Table 4.10 Classification Performance (F-Measure, G-Means and MCC) on 16 Benchmark Dataset using Random Forest (based on Weka) 78

Table 4.11 Classification Performance (F-Measure) of DECIML 79

Table 4.12 Classification Performance (G-Means) of DECIML 79

Table 4.13 Classification Performance (MCC) of DECIML 80

Table 4.14 Win and draw for F-measure of AdaboostM1, Bagging and NB+1NN 81

Table 4.15 Win and draw for F-measure of AdaboostM1, Bagging and NB+kNN 82

Table 4.16 Overall average classification performance of 9 ensemble algorithms on 16 benchmark dataset 83

Table 4.17 Classification accuracy of DECIML ensembles (NB+1NN and NB+kNN) compared to past result from other work 85

Table 4.18 Dataset with low performance of DECIML (#S- Number of Sample, #A-Number of Attribute, #C-Number of Class, #R-Imbalance Ratio) 89

Table 4.19 Sampling with Replacement for DECIML experiment method 91

Table 4.20 F-measure comparison of the DECIML implementation 92

Table 4.21 G-means comparison of the DECIML implementation 93

Table 4.22 MCC comparison of the DECIML implementation 94

Table 4.23 Average values of three evaluation metrics on different ensemble strategies 95

Table 5.1 The taxonomy of feature selection techniques 102

Table 5.2 Information gain (IG) threshold initialization 106

Table 5.3 Algorithm for Multiple Splitting on Attribute Values 107

Table 5.4 GA process for feature selection 109

Table 5.5	Algorithm for population initialization	111
Table 5.6	Imbalance data for feature selection (training and testing set initialization)	113
Table 5.7	Detailed description of the imbalance data for feature selection	113
Table 5.8	Evaluation metric (F-measure (F-m), MCC) and classification accuracies (in %) of DECIML and other selected classifiers using all attributes over the four benchmark dataset	116
Table 5.9	Evaluation metric (F-measure (F-m), MCC) and classification accuracies (in %) of DECIML and other selected classifiers using all attributes over the four benchmark dataset (NB, 1NN and kNN)	117
Table 5.10	Information gain for the dataset	118
Table 5.11	NB+1NN evaluation metric (F-measure (F-m), MCC), classification performance (in percentage) and reduction of attribute (#Att) using DECIMLFS.FIG with threshold ig (IG1 and IG2)	118
Table 5.12	NB+kNN evaluation metric (F-measure (F-m), MCC), classification performance (in percentage) and reduction of attribute (#Att) using DECIMLFS.FIG with threshold ig	119
Table 5.13	Classification performance and reduction of attribute (#Att.FS) using Wrapper based feature selection (DECIMLFS.WR and DECIMLFS.WIG)	120
Table 5.14	Description of dataset used in the experiments and relevant attributes selected by each feature selection method	121
Table 5.15	Classification performance obtained using different classifiers	122
Table 6.1	Selected leaf species for the experimental data	133
Table 6.2	Example of the leaf sample of each species class	134
Table 6.3	Information of the experimental leaf data using (full and half leaf shape features)	136
Table 6.4	Comparison of F-measure (F-m), MCC and	137

	classification performance (%) for CfsSubsetEval on Full and Half leaf features	
Table 6.5	Comparison of F-measure (F-m), MCC and classification performance (%) for DECIMLFS.FIG on Full and Half leaf features	137
Table 6.6	Comparison of F-measure (F-m), MCC and classification performance (%) for DECIMLFS.WIG on Full and Half leaf features	138
Table 6.7	Comparison of F-measure (F-m), MCC and classification performance (%) for DECIMLFS.WR on Full and Half leaf features	138
Table 6.8	Comparison of different feature selection methods over different classifiers using full leaf features	139
Table 6.9	Comparison of different feature selection methods over different classifiers using half leaf feature	140
Table 7.1	The summary of the objectives and findings	147

ABBREVIATION

1NN	1-Nearest Neighbor
ABC	Adaptive Base Class
ABKD	Agent Based Knowledge Discovery
ACM	Association for Computing Machinery
AdaBoost	Adaptive Boosting
AdaBoostM1	Adaptive Boosting Multiclass1
Adacost	Adaptive Boosting with Cost (Misclassification cost-sensitive boosting)
ANN	Artificial Neural Networks
AUC	Area Under Curve
AVA	All-Versus-All
BABoost	Balanced AdaBoost
BHS	Binary Hierarchical Classifier
C4.5	C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan
C5.0	Commercial version of C4.5 (algorithm to generate a decision tree)
CATCH	Canadian Assessment of Tomography for Childhood Head Injury
CNNDM	Class Nearest Neighbor Distance Matrix
CT	Computed Tomography
CWW	Class Confidence Weight
DAG	Directed Acyclic Graph
DataBoost	Data Boosting
DataBoost-IM	Data Sets with Boosting and Data Generation – Imbalance

DB2	Divide-by-2
DBEG	Distribution-Based Example Generation
DBKNN	Density-based-EKNN
DDAG	Decision Directed Acyclic Graph
DECIML	Direct Ensemble Classifier for Imbalanced Multiclass Learning
DECIMLFS	Direct Ensemble Classifier for Imbalanced Multiclass Learning with Feature Selection
DECIMLFS.FIG	Direct Ensemble Classifier for Imbalanced Multiclass Learning with Filter-based feature selection and Information Gain threshold
DECIMLFS.WIG	Direct Ensemble Classifier for Imbalanced Multiclass Learning with Wrapper-based feature selection and Information Gain threshold
DECIMLFS.WR	Direct Ensemble Classifier for Imbalanced Multiclass Learning with Wrapper-based feature selection and Random Information Gain threshold
DS	Decision Stump
DT	Decision Tree
ECOC	Error-Correcting Output Code
eKISS	ensemble Knowledge for Imbalance Sample Sets
EKNN	Evidence-theory-based-KNN
F-measure	F1 Score/Balance F-Score (to measure test accuracy)
FN	False Negative
FP	False Positive
FS	Feature Selection
FSMC	Feature Selection for Minority Class
GA	Genetic Algorithm
GC	Generalized Coding

G-mean	Geometric mean
HSVM	Hierarchical Support Vector Machines
IB1	Instance Based (learning algorithm in Weka)
ID3	Iterative Dichotomiser 3 - is an algorithm used to generate a decision tree invented by Ross Quinlan
IEEE	Institute of Electrical and Electronics Engineers
IG	Information Gain
IR	Imbalance Ratio
J48	Open source Java implementation of the C4.5 algorithm in Weka
KDD	Knowledge Discovery in Databases
KEEL	Knowledge Extraction based on Evolutionary Learning (data repository)
KNN	k-Nearest Neighbor
LI	Lack of Information
LogitBoost	Logistic Boosting
LogitBoost-J	Logistic Boosting (extended for unbalanced data situation)
M1	Model 1
M1v	Model 1 vote
M2	Model 2
M2v	Model 2 vote
MAP	Maximum A-Posteriori
MDLP	Minimum Description Length Principle
MGM	Maximum Geometry Mean
MLNN	Multiclass Leveraged k-Nearest Neighbor
MLP	Multi Layer Perceptron
MLP	Multilayer Perceptron
MMC	Moving Median Center hypersphere

MPEG-7	Multimedia Content Description Interface – 7
MS	Maximum Sum
NB	Naïve Bayes
NIPS	Neural Information Processing Systems Conference (data repository for feature selection challenge)
OAA	One-Against-All
OAO	One-Against-One
OR	Or truth
OVA	One-Versus-All
PAC	Probably Approximately Correct
PAQ	P-Against-Q
PART	Projective Adaptive Resonance Theory
ROC	Receiver Operating Characteristic
RSM	Random Subspace Method
RUSBoost	Random Under-Sampling with Boosting
S2N	Signal to Noise correlation coefficient
SLIPPER	Simple Learner with Iterative Pruning to Produce Error Reduction
SMO	Sequential Minimal Optimization (algorithm)
SMOTE	Synthetic Minority Over-sampling Technique
SMOTEBoost	Synthetic Minority Over-sampling Technique with Boosting
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UCI	University of California, Irvine

UCLA	University of California, Los Angeles
UCR	University of California, Riverside (data repository)
Weka	Waikato Environment for Knowledge Analysis
WINNOW	Machine learning algorithm for learning a linear classifier from labeled examples similar to the perceptron algorithm
fMRI	functional Magnetic Resonance Imaging

SYMBOL

$\subseteq$	Is contained in
$\mathbb{R}^n$	Set of n real number
$\Re^n$	Set of n real number
$\mathbb{H}$	The (set of) quaternions/hypotheses
$\in$	Is an element of
$\|\|$	Parallel; Is parallel to
$\|$	Conditional probability; given
Π	Product; product of all values in range of series
$\sqrt{x}$	Square root
Σ	Summation; sum of all values in range of series
$\hat{f}(x)$	Circumflex/estimator; of the function of x
$\delta(x)$	Delta function
$\delta(x) = \begin{cases} x \\ y \end{cases}$	Dirac delta function; hyperfunction
$=$	Is equal to
$==$	Equivalence
$\geq$	Greater or equal
$>$	Greater than
β	Beta
μ	Mu
$\leftarrow$	Arrow function; from..to; set theory
θ	Theta
Δ	Symmetric difference
ε	Epsilon; represent small number near zero

CHAPTER 1

INTRODUCTION

1.1 Background

Over the decades, knowledge discovery through data mining and machine learning have been extensively studied and applied in various fields. It continues to solve many real world problems and applications, such as pattern recognition, computer vision, image processing, bioinformatics, and a lot more complicated domains. Researches in this massive artificial intelligence domain bring notable advantages where data mining and machine learning algorithms provide assistance in helping people for the knowledge discovery in databases (KDD).

The classic definition of KDD is "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad *et al.*, 1996). Knowledge discovery is one of artificial intelligence contributions to research community which includes the processes of gathering data and information, pre-processing, analyzing data, extracting hidden knowledge (with data mining), constructing knowledge, knowledge evaluations and knowledge reuse. Knowledge discovery is not only using artificial intelligence techniques but it is also supported by other theories in database, retrieval theories, computing, visualization, statistics, etc. Generally, the data mining processes are grouped into three major key steps: preparation of input data, mining of data, and post-processing of output patterns (Du, 2010).

Knowledge and decisions applied today were learned from past experiences and they are iteratively refined to be applied in future problems encountered. Data mining is one of the steps in the knowledge discovery process that can be used to automate the discovery of patterns or data modeling for selected empirical data, visualize and finally use the learned knowledge in response to future unseen data. Data mining is defined as the extraction of implicit, previously unknown, and potentially useful knowledge from data (Witten and Frank, 2000). In other words, data mining can be interpreted as the task of employing an algorithm that processes raw data automatically or semi-automatic and extracts any meaningful patterns that will be used for prediction tasks on new unseen data. Machine learning specifically provides various methods and learning algorithms that can be used to find and describe any structural patterns in data. Thus, the study of machine learning algorithms has emerged as the technical basis for any data mining works.

Currently, many of these common algorithms and their advanced variations provide high classification performance in various empirical data. Advanced techniques, such as ensemble learning methods, are employed which apply different learning algorithms for different applications and these methods provide even higher classification accuracy. For example, ensemble learning approaches have been applied in a shape classification task is able to classify MPEG-7 shape and Swedish leaf shape dataset as high as 95 percent and 98 percent (Temlyakov *et al.*, 2010). As a result, it is substantially harder for new researchers propose any better classification methods. The result of high performance accuracies produced by these advanced techniques indicates that the advancement of data mining methods and machine learning algorithms are almost stagnant and it can solve almost any classification tasks. However, real world problem is far from the reality of no unsolvable classification problem. There is still exist diverse problem requiring different efforts to find efficient solution for comparison, such as new dataset (even in the similar domain which was solved before), massive data, incomplete data (caused by noise or missing values), etc. For example, one of the most challenging data mining problems that are still receiving attention among researchers is the multiclass and imbalance classification problem (Alejo *et al.*, 2008; Ghanem *et al.*, 2010; Lerteerawong and Athimethphat, 2011; Tahir *et al.*, 2010; Valizadegan *et al.*, 2008; Zhou and Liu, 2010).

In data mining, multiclass classification problem refers to assigning one of the several class labels to an input object. Unlike the binary classification, learning a multiclass problem is a more complex task due to the fact that each example can only be assigned to exactly one class label (Valizadegan *et al.*, 2008). In fact, numerous attempts of using binary classification methods have failed to perform