Dedication

To my great educator

Alsheikh Mohamed Mohana Basheir

To my lovely Mother, Father, Sisters, Brothers and Husband.

ACKNOWLEDGEMENTS

Gratefully thanks to my supervisor Prof. Ajith Abraham for his supervision, valuable comments and support throughout the research period.

Table of Contents

Title	Page
Dedication	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	vii
List of Figures	x
List of Abbreviations	xii
Abstract	xiii
المستخلص	xiv
CHAPTER ONE: INTRODUCTION	
1.1 Introduction	1
1.2 Problem Statement	2
1.3 Objectives	4
1.4 Research Questions	4
1.5 Scope	5
1.6 Thesis Structure	5
CHAPTER TWO: Literature Review	
2.1 Imbalanced Two Class classification Review	6
2.1.1 Sampling based methods	6
2.1.1.1 Undersampling	6
2.1.1.2 Oversampling	8
2.1.3 Cost sensitive learning based methods	11
2.1.4 Recognition based methods	13
2.1.5 Ensemble- based Methods	14
2.2 Imbalanced Multi Class classification Review	18

2.3 Summary	20
CHAPTER THREE: EXPERIMENTAL METHODOLOGIES	
3.1 Datasets	21
3.1.1 Two Class Imbalanced Data	21
3.1.2 Multi Class Imbalanced Data	22
3.2. Sampling Methods	24
3.3 The Basic Learners	25
3.4 Meta learning ensembles methods	29
3.4.1 Bagging	29
3.4.2 Boosting	30
3.5. Evaluation Metrics	32
3.6 Summary	33
CHAPTER FOUR: HANDLING TWO CLASS IMBALANCE PROBLEM	
4.1 Experiments Design Methodology	34
4.1.1 Phase One: Testing Classifiers Using the Original Data Distribution	34
4.1.2 Phase Two: Using Resampling Methods	35
4.1.3 Phase Three: Using Meta Learning Methods	35
4.1.4 Phase Four: Using the proposed Approach	35
4.2 Results Analysis and Discussion	37
4.2.1 Results Analysis and Discussion for Phase One	37
4.2.2 Results Analysis and Discussion for Phase Two	42
4.2.3 Results Analysis and Discussion for Phase Three	48
4.2.4 Results Analysis and Discussion for Phase Four	52
CHAPTER FIVE: HANDLING MULTI CLASS IMBALANCED PROBLEM	
5.1 Experiments Design	54
5.1.1 Phase One: Testing Classifiers Using the Original Data Distribution	54

5.1.2 Phase Two: Using Meta Learning Methods	55
5.1.3 Phase Three: The proposed Approach Methodology	55
5.2 Results Analysis and Discussion	
5.2.1 Results Analysis and Discussion for Phase One	59
5.2.2 Results Analysis and Discussion for Phase Two	62
5.2.3 Results Analysis and Discussion for Phase Three	67
5.3. Summary	73
CHAPTER SIX: CONCLUSIONS	
6.1 Conclusions	74
6.2 Future works	75
APPENDICES	77
References	103

List of Tables

Title	Page
Table 2.1. The advantages and drawbacks of the proposed methods for dealing with imbalance problem	17
Table 3.1: A 2x2 Confusion Matrix	32
Table 4.1. Datasets summary	37
Table 4.2. Performance of classifiers on different datasets in term of accuracy	37
Table 4.3 Performance of different classifiers on Insurance Fraud data set using the original data distribution	38
Table 4.4. Performance of different classifiers on German data set using the original data distribution	38
Table 4.5. Performance of different classifiers on Hepatitis data set using the original data distribution	39
Table 4.6. Performance of different classifiers on Haberman data set using the original data distribution	39
Table 4.7. Performance of classifiers on Insurance Fraud data set using undersampling	42
Table 4.8. Performance of different classifiers on German data set using undersampling	43
Table 4.9. Performance of classifiers on Hepatitis data set using undersampling	43
Table 4.10. Performance of classifiers on Haberman data set using undersampling	44
Table 4.11. Performance of classifiers on Insurance Fraud data set using oversampling	45
Table 4.12. Performance of classifiers on German data set using oversampling	46

Table 4.13. Performance of classifiers on Hepatitis data set using oversampling	46
Table 4.14. Performance of classifiers on Haberman data set using oversampling	47
Table 4.15. Performance of classifiers when using meta Learning methods on Insurance Fraud data set	48
Table 4.16. Performance of classifiers when using meta Learning methods on German data sets	49
Table 4.17. Performance of classifiers when using meta Learning methods on Hepatitis data sets	50
Table 4.18. Performance of classifiers when using meta Learning methods on Haberman data sets	51
Table 4.19. Performance of the proposed method on different data sets	53
Table 5.1. One per Class coding	56
Table 5.2. Distributed output coding	56
Table 5.3. Dataset summary	58
Table 5.4. Performance of classifiers on different datasets in term of accuracy	59
Table 5.5. Detection rates per class in Intrusion Detection data set	59
Table 5.6. Detection rates per class in Thyroid data set	60
Table5.7. Detection rates per class in Landsat data set	61
Table 5.8. Detection rates per class in Lymphography data set	61
Table 5.9. The detection rates per class in Glass data set	62
Table 5.10. Detection rates when using Bagging in Intrusion Detection data set	62
Table 5.11. Detection rates when using Bagging in Thyroid data set	63
Table 5.12. Detection rates when using Bagging in Landsat data set	63
Table 5.13. Detection rates when using Bagging in Lymphography data set	63
Table 5.14. Detection rates when using Bagging in Glass data set	64
Table 5.15. Detection rates when using AdaBoost in Intrusion detection data set	64
Table 5.16. Detection rates when using AdaBoost in Thyroid data set	64
Table 5.17. Detection rates when using AdaBoost in Lymphography data set	65

Table 5.18. Detection rates when using AdaBoost in Landsat data set	65
Table 5.19. Detection rates when using AdaBoost in Glass data set	66
Table5.20. Detection rates per class in Intrusion Detection data set using the hybrid proposed ECOC ensemble	70
Table5.21. Detection rates per class in Thyroid data set using the proposed hybrid ECOC ensemble	71
Table 5.22. Detection rates per class in Landsat data set using the proposed hybrid ECOC ensemble	71
Table5.23. Detection rates per class in Lymphography data set using the proposed hybrid ECOC ensemble	72
Table5.24. Detection rates per class in Glass data set using the proposed hybrid ECOC ensemble	72

List of Figures

Title	Page
Figure 1.1: concept complexity (class overlapping) in imbalanced data	3
Figure 1.2: Small disjuncts in imbalanced data	4
Figure 3.1 Algorithm of SMOTE	24
Figure 3.2 Three Layers Back Propagation Neural Networks	26
Fig. 3.3 Radial basis function neural network	27
Figure 3.4 Bagging Algorithm	30
Figure 3.5 AdaBoost Algorithm	31
Figure 4.1 Algorithm of the proposed method for two class problem	36
Figure 4.2 Detection rates for positive and negative classes in Insurance fraud data set	40
Figure 4.3 Detection rates for positive and negative classes in German data set	41
Figure 4.4 Detection rates for positive and negative classes in Hepatitis data set	41
Figure 4.5 Detection rates for positive and negative classes in Haperman data set	42
Figure 4.6. Detection rates of negative and positive classesof fraud data set when using undersampling	44
Figure 4.7. Detection rates of negative and positive classes of insurance fraud data set when using oversampling	47
Figure 5.1 The Pseudo code of the proposed method for multi class problem	57
Figure 5.2 Detection rates per class in Intrusion Detection dataset	67
Figure 5.3 Detection rates per class in Thyroid dataset	67
Figure 5.4 Detection rates per class in Land sat dataset	68
Figure 5.5 Detection rates per class in Lymphography dataset	68

Figure 5.6 Detection rates per class in Glass dataset	69

List of Abbreviations

SMOTE	Synthetic Minority Oversampling Technique
SWR	Sampling with replacement
DECIML	Direct Ensemble Classifier for Iimbalanced Multi Class Learning
NB	Naïve Bays
SVM	Support Vector Machine
BP	Back Propagation Neural Networks
RBF	Radial Basis Function Network
RF	Random Forest
RT	Random Tree
TPR	True Positive Rate
TNR	True Negative Rate
Prec	Precision
F-M	F-Measure
ECOC	Error Correcting Output Code
CNN	Condensed Nearest Neighbor
1NN	One- Nearest Neighbour
OSS	One Sided Selection
GSVM-RU	Granular support vector machines repetitive under-sampling
PSO	Particle Swarm Optimization
C-MEIN	Clustering with Sampling for Multi class Imbalanced using Ensemble
OAA	One-Against-All
OAO	one against one
OAA-DB	One-Against-All with Data Balancing
SS	Sample Selection
OAHO	One Against Higher Order
IR	Imbalance Ratio

Abstract

Class imbalance is one of the challenges of machine learning and data mining fields. Imbalanced data set degrades the performance of data mining and machine learning techniques as the overall accuracy and decision-making would be biased to the majority class, which leads to misclassifying the minority class samples or furthermore treated them as noise. The classification problem of imbalanced data gets complicated whenever the class of interest is relatively rare and has small number of instances compared to the majority class. Moreover, the cost of misclassifying the minority class is very high in comparison with the cost of misclassifying the majority class as occurs in many real applications such as medical diagnosis, fraud detection, network intrusion detection…etc.

In this dissertation, we started by investigating the problem of two class classification. A series of experiments are conducted using imbalanced data with its original distribution, balanced data using sampling methods and meta learning methods. Then, we developed a hybrid ensemble that implemented multi resampling methods at various rates. The experimental results on many real world applications for two class imbalanced data sets, confirms that the proposed hybrid ensembles have better performance using different evaluation measures.

Next, we investigated the multi class imbalanced problem. A series of experiments are conducted using direct multi class classification and meta learning methods. We developed a hybrid Error Correcting Output Code ensemble utilizing weighted Hamming distance and AdaBoost meta learning method. The experimental results on many real applications multi class imbalanced data sets show that our proposed hybrid ensemble performed effectively better by improving the classification performance in minority classes and significantly outperformed other tested methods.

المستخلص

عدم توازن الأصناف هي واحدة من التحديات التي تواجه مجالات تعليم الآلة وتعدين البيانات. مجموعات البيانات التي تحتوي على أصناف غير متوازنة تؤدي إلى تدهور أداء خوارزميات تعليم الاله وتعدين البيانات حيث أن الدقة الكلية وصنع القرار يكون متحيزا للاصناف الأغلبية مما سيؤدي إلى خطأ تصنيف بيانات الأصناف الأقلية أو كحد أقصى التعامل معها كتشويش. تتعقد مشكلة تصنيف البيانات الغير متوازنة متى ما كان الصنف المراد نادراً أو يحتوي على بيانات أقل مقارنة بصنف الأغلبية وأكثر من ذلك عندما تكون تكلفة خطأ التصنيف لصنف الأقلية مرتفع كما نجده في بعض التطبيقات الواقعية مثل التشخيص الطبي والكشف عن حالات الغش وكشف اختراق الشبكة.

في هذا البحث بدأنا بدراسة مشكلة تصنيف البيانات ذات الصنفين، وأجريت عدد من التجارب باستخدام التوزيع الأصلي للبيانات وبيانات متوازنة الأصناف باستخدام تقنيات موازنة عينات البيانات وطرق تعليم المجاميع المتجانسة. ثم قمنا بتطوير طريقة باستخدام المجاميع الهجينة التي تطبق طرق متعددة لموازنة البيانات بمعدلات مختلفة. النتائج التجريبية في عدد من مجموعات البيانات الغير متوازنة لتطبيقات واقعية أكدت على طريقة المجاميع الهجينة المقترحة للحل ذات أداء جيد باستخدام مقاييس تقييم مختلفة.

ثانيا درسنا مشكلة الأصناف المتعددة غير المتوازنة، وأجريت عدد من التجارب باستخدام التصنيف المتعدد المباشر واستخدام طرق تعليم المجاميع المتجانسة. ثم طورنا هجين مجاميع يستخدم طريقة تصحيح خطأ الرمز الناتج وطريقة تعليم المجاميع المتجانسة AdaBoost وطريقة بعد Hamming الموزونة. النتائج التجريبية في عدد من مجموعات البيانات المتعددة الأصناف والغير متوازنة لتطبيقات واقعية أظهرت أن المجاميع الهجينة المقترحة للحل ذات أداء أفضل وفعال من خلال تحسين الأداء للأصناف الأقلية يتفوق على كل الطرق المختبرة الأخرى.