Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

### Complexity Measurement for Dealing with Class Imbalance Problems in Classification Modelling

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy Massey University, 2012

#### Muhammad Nafees Anwar

Institute of Fundamental Sciences Massey University New Zealand To my beloved late Mother

### Abstract

The class imbalance problem is a challenge in the statistical, machine learning and data mining domains. Examples include fraud/intrusion detection, medical diagnosis/monitoring, bioinformatics, text categorization, insurance claims, and target marketing. The problem with imbalanced data sets is that the conventional classifiers (both statistical and machine learning algorithms) aim at maximizing overall accuracy, which is often achieved by allocating all, or almost all, cases to the majority class. Thus there tends to be bias against the minority class in class imbalance situations.

Despite numerous algorithms and re-sampling techniques proposed in the last few decades to tackle imbalanced classification problems, there is no consistent winning strategy for all data sets (neither in terms of sampling, nor learning algorithm). Special attention needs to be paid to the data in hand. In doing so, one should take into account several factors simultaneously: the imbalance rate, the data complexity, the algorithms and their associated parameters. As suggested in the literature, mining such datasets can only be improved by algorithms tailored to data characteristics; therefore it is important and necessary to do data exploratory analysis before deciding on a learning algorithm or re-sampling techniques.

In this study, we have developed a framework "Complexity Measurement" (CM) to explore the connection between the imbalanced data problem and data complexity. Our study shows that CM is an ideal candidate to be recognized as a "goodness criterion" for various classifiers, re-sampling and feature selection techniques in the class imbalance framework. We have used CM as a meta-learner to choose the classifier and under-sampling strategy that best fits the situation. We design a systematic over-sampling technique, Over-sampling using Complexity Measurement (OSCM) for dealing with class overlap. Using OSCM, we do not need to search for an optimal class distribution in order to get favorable accuracy for the minority class, since the amount of over-sampling is determined by the complexity; ideally using CM would detect fine structural differences (class-overlap and small disjunct) between different classes.

Existing feature selection techniques were never meant for class imbalanced data. We propose Feature Selection using Complexity Measurement (FSCM), which can specifically focus on the minority class, hence those features (and multivariate interactions between predictors) can be selected, which form a better model for the minority class.

Methods developed have been applied to real datasets. The results from imbalanced datasets show that CM, OSCM and FSCM are effective as a systematic way of correcting class imbalance/overlap and improving classifier performance. Highly predictive models were built; discriminating patterns were discovered, and automated optimization was proposed. The methodology proposed and knowledge discovered will benefit exploratory data analysis for imbalanced datasets. It may be taken as a judging criterion for new algorithms and re-sampling techniques. Moreover, new data sets may be evaluated using our CM criterion in order to build a sensible model.

vi

## Acknowledgements

I am highly in debt to Dr. Geoff Jones for his invaluable guidance and help to me in the development of my research skills, and his efforts to keep a stray bird on the right track.

I am very grateful to Dr. Siva Ganesh. Ganesh has always allowed me complete freedom to define and explore my own directions in research. While this proved difficult and somewhat bewildering to begin with, I have come to appreciate the wisdom of his way, as it encouraged me to think for myself.

I am thankful to Dr. Martin Hazelton and the Statistic Group who have provided me with financial support during my degree. They have kindly provided casual teaching assistantship positions and travel funds to attend conferences.

I thank Dr. Ganes Ganesalingam and Dr. Mark Bebbington for providing valuable feedback on parts of this thesis.

I also thank Mr. Peter Lewis for providing technical support.

Special thanks must also go to my family, especially to my late mother and to my wife Wa. They have provided unconditional support and encouragement through both the highs and lows of my time in Ph.D study.

Finally I would like to thank Higher Education Commission (HEC), Pakistan, for providing me with a scholarship to study at Massey University, New Zealand.

# Contents

A	bstra	$\mathbf{ct}$		iii
A	cknov	wledge	ements	vii
$\mathbf{Li}$	st of	Figur	es	xv
$\mathbf{Li}$	st of	Table	S	xix
1	Intr	oducti	ion	1
	1.1	Appro	baches for Dealing with Class Imbalance	3
		1.1.1	Limitations of Traditional Learning Techniques	4
		1.1.2	Pre-processing Techniques	4
	1.2	Comp	lexity Measurement Approach	8
		1.2.1	Overview of Literature Review (Chapter 2)	8
		1.2.2	Overview of Complexity Measurement (Chapter 3) $\therefore$	9
		1.2.3	Overview of Under-sampling Techniques in relation to	
			Complexity Measurement (Chapter 4)	9
		1.2.4	Overview of Over-sampling Techniques using Complex-	
			ity Measurement (Chapter 5)	10
		1.2.5	Overview of Feature Selection using Complexity Mea-	
			surement (Chapter 6) $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$	11
	1.3	Contr	ibution of the Thesis:	12
Re	efere	nces		13
<b>2</b>	Lite	rature	e Review of Class Imbalance Problem	<b>21</b>
	2.1	Introd	luction	22
	2.2	Impor	tance of Class Imbalance Problems:	24
	2.3	Dealir	ng with Class Imbalance Problems	25
		2.3.1	Re-sampling Techniques	25
			2.3.1.1 Under-Sampling Techniques	26
			2.3.1.2 Random over-sampling	28

		2.3.1.3 Active Re-sampling Techniques	9
		2.3.2 Cost-Sensitive Learning	1
		2.3.2.1 Tuning parameter for Classifiers	3
		2.3.2.2 One class learning	4
		2.3.3 Feature Selection	5
		2.3.4 Ensemble Classifiers	6
	2.4	Evaluation Measures	7
		2.4.1 Probabilistic Classifiers	9
		2.4.2 F-measure	0
		2.4.3 G-mean	1
		2.4.4 Receiver Operating Characteristic (ROC) 4	1
	2.5	Discussion $\ldots \ldots 43$	3
	2.6	Conclusion $\ldots \ldots 44$	4
B	efere	nces d'	7
10	2.7	Appendix "Classification Methods"	6
	2.1	271 Logistic regression	6
		$2.7.1$ Logistic regression $\dots \dots \dots$	6
		2.7.2 Chassification frees	7
		2.7.4 Support vector machines	8
			0
3	Me	asurement and Visualization of Data Complexity for Clas-	-
3	Mea sific	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data 6	1
3	Mea sific 3.1	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data 6 Introduction	1
3	Mea sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         Output       6	<b>1</b> 1 3
3	Mea sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data6Introduction6Data Complexity63.2.1Studies on Data Complexity6	<b>1</b> 1 3 4
3	Mea sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data6Introduction6Data Complexity63.2.1Studies on Data Complexity63.2.2Data Complexity Measures6	<b>1</b> 3 4 4
3	Mea sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data6Introduction6Data Complexity63.2.1Studies on Data Complexity63.2.2Data Complexity Measures63.2.3Complexity, Overlap and Imbalance6	$     \begin{array}{c}       1 \\       51 \\       3 \\       4 \\       4 \\       7 \\       \end{array} $
3	Mea sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data6Introduction6Data Complexity63.2.1Studies on Data Complexity63.2.2Data Complexity Measures63.2.3Complexity, Overlap and Imbalance63.2.4Bayes Error6	$1 \\ 3 \\ 4 \\ 7 \\ 8 \\ 0$
3	Mes sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6	$1 \\ 1 \\ 3 \\ 4 \\ 7 \\ 8 \\ 9 \\ 1$
3	Me: sific 3.1 3.2	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7	$     \begin{array}{c}       1 \\       1 \\       3 \\       4 \\       4 \\       7 \\       8 \\       9 \\       1     \end{array} $
3	Mes sific 3.1 3.2 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7	$1 \\ 1 \\ 3 \\ 4 \\ 4 \\ 7 \\ 8 \\ 9 \\ 1 \\ 1$
3	Me: sific 3.1 3.2 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Nearest Neighbors       7	1 1 3 4 4 7 8 9 1 1 2
3	Mea sific 3.1 3.2 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         3.3.1       Distance metric       7	113447891 135
3	Me: sific 3.1 3.2 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         3.3.1       Distance metric       7         3.3.2       Visualization       7	1134478911352
3	Me: sific 3.1 3.2 3.3 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Nearest Neighbors       7       7         3.3.1       Distance metric       7         Simulation Study       7       7	<b>1</b> 13 44 78 91 13 56 6
3	Me: sific 3.1 3.2 3.3 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Nearest Neighbors       7       3.3.1       Distance metric       7         Simulation Study       7       3.4.1       Design       7	$1 \\ 1 \\ 3 \\ 4 \\ 4 \\ 7 \\ 8 \\ 9 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1$
3	Me: sific 3.1 3.2 3.3 3.3	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Sa.1       Distance metric       7         3.3.1       Distance metric       7         Simulation Study       7       7         3.4.1       Design       7         3.4.2       Results       7	$1 \\ 1 \\ 3 \\ 4 \\ 4 \\ 7 \\ 8 \\ 9 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 1 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1$
3	Me: sific 3.1 3.2 3.3 3.3 3.4 3.5	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Nearest Neighbors       7       7         3.3.1       Distance metric       7         Simulation Study       7       7         3.4.1       Design       7         3.4.2       Results       7         3.4.3       Distance       7         3.4.4       Design       7         3.4.5       Results       7	113447891 $1356671$
3	Me: sific 3.1 3.2 3.3 3.3 3.4 3.5	asurement and Visualization of Data Complexity for Clas- cation Problems with Imbalanced Data       6         Introduction       6         Data Complexity       6         3.2.1       Studies on Data Complexity       6         3.2.2       Data Complexity Measures       6         3.2.3       Complexity, Overlap and Imbalance       6         3.2.4       Bayes Error       6         3.2.4.1       Parametric Estimate of the Bayes Error       6         3.2.4.2       Non-Parametric Estimates of the Bayes Error       7         Methodology:       Data Complexity Measurement based on K-       7         Simulation Study       7       7         3.4.1       Design       7         3.2.2       Visualization       7         3.3.1       Distance metric       7         3.4.1       Design       7         3.4.2       Results       7         3.4.2       Results       7         3.4.2       Results       7         3.5.0.1       Threshold to regulate k in Complexity Mea-	<b>1</b> <b>1</b> <b>3</b> <b>4</b> <b>4</b> <b>7</b> <b>8</b> <b>9</b> <b>1</b> <b>1</b> <b>3</b> <b>5</b> <b>6</b> <b>6</b> <b>7</b> <b>1</b>

\_\_\_\_\_

	3.6	Discus	ssion and Future Work	39
Re	efere	nces	9	3
	3.7	Apper	ndix	)7
		3.7.1	R Codes	97
		3.7.2	Results for 3 Dimensions:	)8
4	An	Empir	ical Comparison of Under-Sampling Techniques in	
	$\operatorname{Rel}$	ation t	to Data Complexity 10	)1
	4.1	Introd	luction $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ 10	)2
	4.2	Curre	nt Under-Sampling (US) Approaches	)3
		4.2.1	Random under-sampling (RUS):	)5
		4.2.2	Neighborhood Cleansing Techniques (NCT) 10	)5
			4.2.2.1 Tomek links:	)5
			4.2.2.2 Condensed Nearest Neighbor Rule: 10	)6
			4.2.2.3 One-sided selection $(OSS)$ :	)6
			4.2.2.4 Neighborhood Cleaning Rule (NCL): 10	)7
		4.2.3	Active Learning	)7
			4.2.3.1 Progressive Learning:	)7
		4.2.4	Repetitive Under-Sampling	)8
			4.2.4.1 EasyEnsemble:	)8
			4.2.4.2 RUSBoost:	)8
			4.2.4.3 Classification using lOcal clusterinG (COG): . 10	)8
		4.2.5	Cluster-based Under-sampling	)9
			4.2.5.1 Clustering Undersampling:	)9
	4.3	Empir	rical Study:	0
		4.3.1	Learning algorithms	11
			4.3.1.1 Ensemble classification methods	2
			$4.3.1.2  \text{Clustering}  \dots  \dots  \dots  \dots  \dots  \dots  \dots  \dots  \dots  $	3
		4.3.2	Study Design	3
		4.3.3	Statistical Analysis: Friedman Test	5
		4.3.4	Results:	6
	4.4	Discus	ssion	30
		4.4.1	Validity of Results:	31
	4.5	Concl	usion $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $13$	32
Re	efere	nces	13	5
	4.6	Apper	$dix \ldots 14$	11
		4.6.1	R codes $\ldots \ldots 14$	11

of Minority Class in Imbalanced Data Sets       147         5.1       Introduction       147         5.2       Over-sampling using Complexity Measure (OSCM)       150         5.2.1       Complexity measure       152         5.2.2       SMOTE       152         5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE:       156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2.1       Results       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166         References         167       5.6         5.6       Appendix       169         5.6.1       R Codes       167         5.6       Appendix       169         5.6.1       R Codes       169         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.1 <td< th=""><th><b>5</b></th><th colspan="5">Complexity Measure: A Systematic Approach to Over-sampling</th></td<>	<b>5</b>	Complexity Measure: A Systematic Approach to Over-sampling				
5.1       Introduction       147         5.2       Over-sampling using Complexity Measure (OSCM)       150         5.2.1       Complexity measure       152         5.2.2       SMOTE       152         5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE:       156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2.1       Results       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166         References       167         5.6       Appendix       169         5.6.1       R Codes       169         5.6.1       R Codes       169         5.6.1       R Codes       173         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection and Lagorithms       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180		of $\mathbb{N}$	/linorit	ty Class in Imbalanced Data Sets	147	
5.2       Over-sampling using Complexity Measure (OSCM)       150         5.2.1       Complexity measure       152         5.2.2       SMOTE       154         5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE: 156       153         Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2.1       Results       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166 <b>References</b> 167         5.6       Appendix       169         5.6.1       R Codes       169         5.6.1       R Codes       169         5.6.1       R Codes       175         6.2       Urrent Feature Selection (FS) Approaches       175         6.3.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.2.2       Random Forest (RF)       182		5.1	Introd	luction	. 147	
5.2.1       Complexity measure       152         5.2.2       SMOTE       152         5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE:       156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2       Real Data Examples       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166 <b>References</b> 167         5.6       Appendix       169         5.6.1       R Codes       169         5.6.1       R Codes       169         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       181         6.3.2.3       Boruta Algorithm       183         6.3.2.4       Renetic Algorithm       183         6.3.2.1       Genetic Algorithm       183		5.2	Over-s	sampling using Complexity Measure (OSCM)	. 150	
5.2.2       SMOTE       152         5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE: 156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       157         5.3.2       Real Data Examples       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166         References         167       5.6         5.6       Appendix       169         5.6.1       R Codes       169         5.6.1       R Codes       169         5.6.1       R Codes       173         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection in Class Imbalance       178         6.3       Feature Selection Algorithms       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.2.3       Boruta Algorithm (GA)       181         6.3.2.3       Boruta Algo			5.2.1	Complexity measure	. 152	
5.2.3       SMOTE using complexity measure: (SCM)       154         5.2.4       Performance measure for over sampling using SMOTE: 156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2.1       Results       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166 <b>References</b> 167         5.6       Appendix       169         5.6.1       R Codes       169         5.6.1       R Codes       169 <b>6 Feature Selection for Classification Problems with Imbalanced Data</b> 173         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection algorithms       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.2.3       Boruta Algorithm (GA)       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.3       Transf			5.2.2	SMOTE	. 152	
5.2.4       Performance measure for over sampling using SMOTE: 156         5.3       Experiments       157         5.3.1       Choice of k       157         5.3.2       Real Data Examples       158         5.3.2.1       Results       160         5.4       Discussion:       164         5.5       Conclusions and Future Work       166 <b>References</b> 167         5.6       Appendix       169         5.6.1       R Codes       169         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection I Class Imbalance       179         6.3.1.2       Entropy based techniques       179         6.3.1.2       Entropy based techniques       180         6.3.2.2       Random Forest (RF)       18			5.2.3	SMOTE using complexity measure: (SCM)	. 154	
5.3Experiments1575.3.1Choice of k1575.3.2Real Data Examples1585.3.2.1Results1605.4Discussion:1645.5Conclusions and Future Work166 <b>References</b> 1675.6Appendix1695.6.1R Codes1696Feature Selection for Classification Problems with ImbalancedData1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.1Chi-square1796.3.1.2Entropy based techniques1806.3.2.3Boruta Algorithm1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection Using CM1866.4.1Heuristic Search1866.4.1.1Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Setz1896.5.1Artificial Data Setz1896.5.2Real Data Setz180			5.2.4	Performance measure for over sampling using SMOTE	: 156	
5.3.1Choice of k1575.3.2Real Data Examples1585.3.2.1Results1605.4Discussion:1645.5Conclusions and Future Work166 <b>References</b> 1675.6Appendix1695.6.1R Codes1695.6.1R Codes1696Feature Selection for Classification Problems with ImbalancedData1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1.1Chi-square1796.3.1.2Entropy based techniques1806.3.2.2Wrapper-based Feature Selection Techniques1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Heuristic Search1866.4.1.1Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Setz1896.5.1Artificial Data Setz1806.5.1Artificial Data Setz1896.5.1Artificial Data Setz1896.5.2Real Data Setz1806.5.1Artificial Data Setz1896.5.2Real Data Setz1806.5.1 <t< td=""><td></td><td>5.3</td><td>Exper</td><td>iments</td><td>. 157</td></t<>		5.3	Exper	iments	. 157	
5.3.2Real Data Examples158 $5.3.2.1$ Results160 $5.4$ Discussion:164 $5.5$ Conclusions and Future Work166 <b>References</b> 167 $5.6$ Appendix169 $5.6.1$ R Codes169 $5.6.1$ R Codes169 $6$ Feature Selection for Classification Problems with ImbalancedData173 $6.1$ Introduction174 $6.2$ Current Feature Selection (FS) Approaches175 $6.2.1$ Feature Selection in Class Imbalance178 $6.3$ Feature selection Algorithms179 $6.3.1.1$ Chi-square179 $6.3.1.2$ Entropy based techniques180 $6.3.2.1$ Genetic Algorithm (GA)181 $6.3.2.1$ Genetic Algorithm183 $6.3.2.3$ Boruta Algorithm183 $6.3.3$ Transformation-based Feature Selection184 $6.4$ Proposed Algorithm185 $6.4.1$ Heuristic Search186 $6.4.1.2$ Sequential Forward Selection Using CM187 $6.5$ Experiments:189 $6.5.1$ Artificial Data Set:189 $6.5.2$ Real Data Set:189 $6.5.4$ Data Set:189 $6.5.4$ Data Set:189 $6.5.4$ Data Set:189 $6.5.4$ Real Data Set:189 $6.5.4$ Real Data Set:189 $6.5.4$ Real Data Set:189 $6.5.4$ <td< td=""><td></td><td></td><td>5.3.1</td><td>Choice of <math>k</math></td><td>. 157</td></td<>			5.3.1	Choice of $k$	. 157	
5.3.2.1Results1605.4Discussion:1645.5Conclusions and Future Work166References1675.6Appendix1695.6.1R Codes1696Feature Selection for Classification Problems with ImbalancedData1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.2Entropy based techniques1806.3.2.3Relief1806.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Heuristic Search1866.4.1.2Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Set:1896.5.2Data Set:1896.5.3Data Set:1896.5.1Arabel Data Set:1896.5.2Data Data Set:1896.5.3Data Data Set:1896.5.4Data Set:1896.5.4Data Set:1896.5.5Aratificial Data Set:1896.5.1Aratificial Data Set:1896.5.2Data Data Set:189			5.3.2	Real Data Examples	. 158	
5.4Discussion:1645.5Conclusions and Future Work166References1675.6Appendix1695.6.1R Codes1696Feature Selection for Classification Problems with ImbalancedData1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.1Chi-square1796.3.1.2Entropy based techniques1806.3.2Wrapper-based Feature Selection Techniques1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Heuristic Search1866.4.1.2Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Set:1896.5.2Real Data Set:180				5.3.2.1 Results	. 160	
5.5Conclusions and Future Work166References1675.6Appendix1695.6.1R Codes1695.6.1R Codes1696Feature Selection for Classification Problems with ImbalancedData1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.2Entropy based techniques1806.3.2Wrapper-based Feature Selection Techniques1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Heuristic Search1866.4.1.2Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Set:1896.5.2Real Data Set:180		5.4	Discus	ssion:	. 164	
References167 $5.6$ Appendix169 $5.6.1$ R Codes169 $6$ Feature Selection for Classification Problems with ImbalancedData173 $6.1$ Introduction174 $6.2$ Current Feature Selection (FS) Approaches175 $6.2.1$ Feature Selection in Class Imbalance178 $6.3$ Feature selection Algorithms179 $6.3.1$ Filter-Based Feature Ranking Techniques179 $6.3.1.2$ Entropy based techniques180 $6.3.2$ Wrapper-based Feature Selection Techniques181 $6.3.2$ Random Forest (RF)182 $6.3.3$ Transformation-based Feature Selection184 $6.4.1$ Heuristic Search186 $6.4.1.1$ Sequential Forward Selection Using CM187 $6.5$ Experiments:189 $6.5.1$ Artificial Data Set:189 $6.5.2$ Read Data Set:189 $6.5.2$ Read Data Set:189		5.5	Conclu	usions and Future Work	. 166	
5.6Appendix1695.6.1R Codes1696Feature Selection for Classification Problems with Imbalanced1736.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.2Entropy based techniques1806.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Sequential Eorward Selection Using CM1866.4.1.2Sequential Backward Selection Using CM1876.5Experiments:1896.5Artificial Data Set:1896.5Artificial Data Set:189	Re	efere	nces		167	
5.6.1       R Codes       169         6       Feature Selection for Classification Problems with Imbalanced Data       173         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection in Class Imbalance       178         6.3       Feature selection Algorithms       179         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.2       Entropy based techniques       180         6.3.2.3       Relief       180         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189		5.6	Appen	ndix	. 169	
6       Feature Selection for Classification Problems with Imbalanced Data       173         6.1       Introduction       174         6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection in Class Imbalance       178         6.3       Feature selection Algorithms       179         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189			5.6.1	R Codes $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$	. 169	
Data173 $6.1$ Introduction174 $6.2$ Current Feature Selection (FS) Approaches175 $6.2.1$ Feature Selection in Class Imbalance178 $6.3$ Feature selection Algorithms179 $6.3.1$ Filter-Based Feature Ranking Techniques179 $6.3.1.1$ Chi-square179 $6.3.1.2$ Entropy based techniques180 $6.3.2$ Wrapper-based Feature Selection Techniques181 $6.3.2.1$ Genetic Algorithm (GA)181 $6.3.2.2$ Random Forest (RF)182 $6.3.2.3$ Boruta Algorithm183 $6.3.4.1$ Genetic Selection Techniques184 $6.4$ Proposed Algorithm185 $6.4.1$ Heuristic Search186 $6.4.1.2$ Sequential Forward Selection Using CM187 $6.5$ Experiments:189 $6.5.1$ Artificial Data Set:189 $6.5.2$ Paral Data Set:180	6	Feat	ture Se	election for Classification Problems with Imbalanc	ed	
6.1Introduction1746.2Current Feature Selection (FS) Approaches1756.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.1Chi-square1796.3.1.2Entropy based techniques1806.3.2Wrapper-based Feature Selection Techniques1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1846.4Proposed Algorithm1856.4.1Heuristic Search1866.4.1.2Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Set:189		Dat	a		173	
6.2       Current Feature Selection (FS) Approaches       175         6.2.1       Feature Selection in Class Imbalance       178         6.3       Feature selection Algorithms       179         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189		6.1	Introd	luction	. 174	
6.2.1Feature Selection in Class Imbalance1786.3Feature selection Algorithms1796.3.1Filter-Based Feature Ranking Techniques1796.3.1.1Chi-square1796.3.1.2Entropy based techniques1806.3.1.3Relief1806.3.2Wrapper-based Feature Selection Techniques1816.3.2.1Genetic Algorithm (GA)1816.3.2.2Random Forest (RF)1826.3.3Transformation-based Feature Selection1836.4.1Heuristic Search1866.4.1.1Sequential Forward Selection Using CM1876.5Experiments:1896.5.1Artificial Data Set:1896.5.2Raal Data Set:189		6.2	Currei	nt Feature Selection (FS) Approaches	. 175	
6.3       Feature selection Algorithms       179         6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.1.3       Relief       180         6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Boal Data Sets       190			6.2.1	Feature Selection in Class Imbalance	. 178	
6.3.1       Filter-Based Feature Ranking Techniques       179         6.3.1.1       Chi-square       179         6.3.1.2       Entropy based techniques       180         6.3.1.3       Relief       180         6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Real Data Sets       190		6.3	Featur	re selection Algorithms	. 179	
6.3.1.1Chi-square179 $6.3.1.2$ Entropy based techniques180 $6.3.1.3$ Relief180 $6.3.2$ Wrapper-based Feature Selection Techniques181 $6.3.2.1$ Genetic Algorithm (GA)181 $6.3.2.2$ Random Forest (RF)182 $6.3.2.3$ Boruta Algorithm183 $6.3.3$ Transformation-based Feature Selection184 $6.4$ Proposed Algorithm185 $6.4.1.1$ Sequential Forward Selection Using CM186 $6.4.1.2$ Sequential Backward Selection Using CM187 $6.5$ Experiments:189 $6.5.1$ Artificial Data Set:180			6.3.1	Filter-Based Feature Ranking Techniques	. 179	
6.3.1.2Entropy based techniques180 $6.3.1.3$ Relief180 $6.3.2$ Wrapper-based Feature Selection Techniques181 $6.3.2.1$ Genetic Algorithm (GA)181 $6.3.2.2$ Random Forest (RF)182 $6.3.2.3$ Boruta Algorithm183 $6.3.3$ Transformation-based Feature Selection184 $6.4$ Proposed Algorithm185 $6.4.1$ Heuristic Search186 $6.4.1.2$ Sequential Forward Selection Using CM187 $6.5$ Experiments:189 $6.5.1$ Artificial Data Set:189 $6.5.2$ Bash Data Sets190			0.0.2	6.3.1.1 Chi-square	. 179	
6.3.1.3       Relief       180         6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       183         6.4.1       Heuristic Search       185         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189				6.3.1.2 Entropy based techniques	. 180	
6.3.2       Wrapper-based Feature Selection Techniques       181         6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       183         6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       180				6.3.1.3 Relief	. 180	
6.3.2.1       Genetic Algorithm (GA)       181         6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       183         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189			6.3.2	Wrapper-based Feature Selection Techniques	. 181	
6.3.2.2       Random Forest (RF)       182         6.3.2.3       Boruta Algorithm       183         6.3.3       Transformation-based Feature Selection       183         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       190			0.0.2	6.3.2.1 Genetic Algorithm (GA)	. 181	
6.3.2.2       Random Foresc (Ref) + + + + + + + + + + + + + + + + + + +				6.3.2.2 Bandom Forest (BF)	182	
6.3.3       Transformation-based Feature Selection       184         6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       185         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       190				6.3.2.3 Boruta Algorithm	183	
6.4       Proposed Algorithm       185         6.4.1       Heuristic Search       185         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       190			633	Transformation-based Feature Selection	184	
6.4.1       Heuristic Search       186         6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       190		64	Propo	sed Algorithm	185	
6.4.1.1       Sequential Forward Selection Using CM       186         6.4.1.2       Sequential Backward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       190		0.1	6 4 1	Heuristic Search	186	
6.4.1.2       Sequential Forward Selection Using CM       187         6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Beal Data Sets       100			0.1.1	6.4.1.1 Sequential Forward Selection Using CM	186	
6.5       Experiments:       189         6.5.1       Artificial Data Set:       189         6.5.2       Boal Data Seta       100				6.4.1.2 Sequential Backward Selection Using CM	187	
6.5.1 Artificial Data Set:		65	Exper	iments.	180	
$6.5.2  \text{Pool Data Seta} \qquad 100$		0.0	651	Artificial Data Set:	180	
			652	Real Data Sets	100	

		6.5.3 Results:
	6.6	Discussion and Future Work:
Re	efere	nces 197
	6.7	Appendix
		6.7.1 R Codes
		6.7.2 Results
7	Disc	cussion 207
	7.1	Discussion and Conclusion of Chapter 3
	7.2	Discussion and Conclusion of Chapter 4
	7.3	Discussion and Conclusion of Chapter 5
	7.4	Discussion and Conclusion of Chapter 6
	7.5	Large data sets
	7.6	Conclusions and Future Work:
Re	efere	nces 221
	7.7	Appendix

## List of Figures

2.1 ROC Curve: the points on the curve represent the performance of the classifier. The ideal model is one that obtains a True Positive Rate of one and a zero False Positive Rate (i.e., TPrate = 1 and FPrate = 0, point A in figure. A worst case scenario would be point B, coordinates (1,0), where TPR is zero and FPR is a maximum. A model that makes a random guess should reside along the line connecting the points (TPrate = 0, FPrate = 0), where every instance is predicted as a negative class, and (TPrate = 1, FPrate = 1), where every instance is predicted as a positive class. . . . . . . . . . . . . . . . .

#### 3.1 Overlap for two classes with a single feature, $x \ldots \ldots \ldots 67$

42

79

- 3.2 Complexity measurement for simulated 3 dimensional normal multivariate distributions. The different panels show different degree of imbalance from the balanced distribution 1000 observation is each class(top left), imbalanced data set 1000 observation in class 1 and 500 in class 2 (top right), imbalanced data set 1000 observation in class 1 and 300 in class 2 (bottom left), to severely imbalanced 1000 observation in class 1 and 90 in class 2 (bottom right). The complexity measure is shown first for Class 1, then for Class 2, for 3-, 5-, 7- and 9-NN with error bars to show variability across simulations. B1, B2 give Bayes error and N1 gives fraction of points on class boundary.
- 3.3 Scatter plot for Sensitivity values and Complexity Measurement for UCI data set, top left is our proposed Complexity Measurement (CM), top right is Fisher Discriminant Ratio (F1), bottom left is Nonlinearity of nearest neighbor or linear classifier (L3) and bottom right shows Fractions of points on class boundary (N1). Every data set is represented by its name 85

3.4	Visualization by MDS of Euclidean distance (left) and Ran- dom forest distances (right) for UCI breast cancer dataset. The symbols and color combination shows two different classes: B (Maroon) benign cases and malignant cases (minority class) shown by the number of 3-nearest-neighbours (Black=3, Pur- ple=2, Blue=1 and Red=Overlapped)	86
3.5	Visualization by MDS of Euclidean distance (left) and Ran- dom forest distances (right) UCI abalone dataset The sym- bols and color combination shows two different classes:G (Red) shows the majority class, where as minority class (Age=7) shown by the number of 3-nearest-neighbours (Black=3, Pur- ple=2, Blue=1 and Red=O)	88
4.1	Scatter plot for Sensitivity values (average sensitivity value from different classifiers) and Complexity Measurement for UCI data set, top left is our proposed Complexity Measure- ment (CM), top right is Imbalance Ratio (IR), bottom left is Fisher Discriminant Ratio (F1) and bottom right shows Frac- tions of points on class boundary (N1). Every data set is represented by its name	117
4.2	A summary of the results obtained by the learning algorithms on the different categories of problems. Each panel shows the results for a different range of complexity (CM) with each classifier evaluated by five different measures along with their 95% confidence interval (shown by an error line at the top of each bar).	121
4.3	A difference obtained by the learning algorithms on the differ- ent data sets for group $CM > 50$ . The result for a differences of AUC values among the classifier evaluated by Friedman Test. On x-axis all possible combination of difference between clas- sifier, where as in legend relative p-values of post hoc test is given	122
4.4	A summary of the results obtained by the learning algorithms on the different categories of problems using the under-sampling technique that gave the best sensitivity along with their 95% confidence interval (shown by error line at top of each bar). The left hand graphs represent the best random under-sampling techniques, whereas right hand shows best NCT techniques.	124

4.5	A summary of the results obtained by the Repetitive undersampling on the different categories of problems, top left represent Complexity Measurement 0% <cm<math>\leq40%, top right 40%<cm<math>\leq50% and bottom centered is CM &gt; 50%. The error bar at the top of each bar represent 95% confidence interval. Every Methodology and Evaluation Measure is represented by its names <math>\therefore</math> 129</cm<math></cm<math>
5.1	Average ROC curve for test set obtained from Pima Indian Diabetic data sets
6.1 6.2	Visualization by One dimension selected by proposed algo- rithm (left) and MDS (right) for UCI Iris data set. The names and color combination shows different classes
<ul><li>7.1</li><li>7.2</li></ul>	Classification for Severe Imbalance distribution. Accuracy of majority class (specificity) is shown by triangles (red) and Ac- curacy of minority class (sensitivity) is shown by the circle (black) over different level of overlap $\ldots \ldots \ldots$

# List of Tables

3.1	Data Complexity Measures by Ho and Basu
3.2	$k$ -Nearest Neighbor Simulation Design $\ldots \ldots \ldots$
3.3	Description of UCI data sets
3.4	This tables compares Sensitivity values and G Mean on UCI Imbalanced data sets. Mean values for each data set were cal- culated for 5 runs with different test subsets obtain from strat- ified 5 fold cross validation The first column lists the data sets used. The following columns(2-3) shows the ratio of minority class in the data set, columns (4-6) the complexity measure used in the literature and column(7) our proposed complexity
	measure
3.5	MB= Mahalanobis Bounds, BLB= Bhattacharyya Lower Bound and BUB= Bhattacharyya Upper Bound 100
4.1	Under-Sampling Techniques
4.2	Description of UCI data sets
4.3	Results for Raw Data Sets
4.4	Best under-sampling technique by Geometric Mean 118
4.5	Best under-sampling technique by Sensitivity
4.6	Best under-sampling technique by AUC
4.7	Results for Repetitive Techniques
5.1	Description of UCI data sets
5.2	Characteristics of the 13 data sets we used in experiment: Na-
	ture of Variables and Class difference. For some data sets the
	class label in the parentheses indicate the target class we chose.
	Moreover, this table shows the choice of optimal over-sampling
	shown by % Over and relative figure show percentage amount
	of oversampling in the minority class for Border line Smote and
	our Complexity Measurement (CM) algorithm with parameter
	$k$ along with resultant amount of over-sampling $\ldots \ldots \ldots 160$

5.3	Sensitivity comparison of methods applied to UCI data sets. Figures gives mean (standard deviation) of sensitivity values calculated using stratified 5 fold cross validation	161
5.4	G-means comparison of methods applied to UCI data sets. Figures gives mean (standard deviation) of sensitivity values	101
5.5	calculated using stratified 5 fold cross validation Over-sampling Effect: True Positive and False Positive of the minority class and Variance of base classifier for True Positive and False Positive	162 165
61	Fasture Selection Techniques	176
6.2	Description of UCI data sets	191
6.3	Sensitivity and AUC comparison of Feature Selection methods applied to WDBC data sets. Rows gives the various Feature Selection methodologies is used. Figures gives mean values	101
6.4	calculated using stratified 5 fold cross validation Sensitivity and AUC comparison of Feature selection methods applied to Yeast data sets. Rows gives the various Feature selection methodologies is used. Figures gives mean values calculated using stratified 5 fold cross validation	191
6.5	Sensitivity and AUC comparison of Feature selection methods applied to highly imbalanced Satimage and Abalone data sets. Rows gives the various Feature selection methodologies is used. Figures gives mean values calculated using stratified 5 fold	102
	cross validation	194
7.1	Confusion Matrices for the Motor Insurance Data set (No Claims: 285,299 (95.1%); Claims: 14,701 (4.9%)) and Forest Cover Type Data set (Spruce-Fir('0'): 211,840 (94.5%);	
	Cottonwood/Willow and Aspen ('1'), $12,240 (5.5\%)$ )	216
7.2	Confusion Matrix for the Motor Insurance Data sets: Roughly Balanced Random Partition (No Claims:15015 (50.5%);Claims:	010
	14,(01 (49.5%))	216