

Table of Contents

Abstract.....	vi
Table of Contents.....	viii
List of Figures	xi
List of Tables.....	xiii
Chapter-1: Introduction.....	1
1.1 Overview of Data Mining.....	1
1.2 The Data Mining Process.....	4
1.2.1 Data Selection.....	5
1.2.2 Data Preprocessing.....	6
1.2.3 Data Mining.....	7
1.2.4 Pattern Evaluation.....	7
1.2.5 Knowledge Presentation	7
1.2.6 Interpretation.....	7
1.2.7 Use of discovered knowledge	7
1.3 Data mining task.....	8
1.3.1 Predictive mining task.....	8
1.3.2 Descriptive task	9
1.4 Models for Data mining Tasks.....	10
1.5 Data Mining Tools.....	11
1.5.1 WEKA	11
1.5.2 ORANGE.....	12
1.6 Problem Statement and Objectives.....	13
1.6.1 Problem statement	13
1.6.2 Research Objectives	13
1.7 Research Contribution.....	13
1.8 Organization of the thesis.....	16
Chapter- 2: Literature Review	18
2.1 Data Mining Techniques.....	18
2.2 Data Mining Tools.....	22
2.3 Improve the Performance of Classification Algorithm.....	30
2.4 Imbalanced Data.....	33
2.5 Feature Selection.....	35
2.6 Research gap.....	37
2.7 Summary.....	38

Chapter-3: Clustering & Classification	39
3.1 Classification	39
3.2 Classifiers Performance measuring Parameters	40
3.3 Classification techniques	42
3.3.1 Regression	44
3.3.2 Bayesian Classification	44
3.3.3 Decision Tree	45
3.3.4 The KNN (K-Nearest Neighbor)	48
3.3.5 Support Vector Machines	48
3.3.6 NN supervised learning	49
3.3.7 Random Forests	51
3.4 Clustering	53
3.4.1 Clustering Techniques	54
3.5 K-means algorithm	55
3.6 Distance metrics	57
3.7 Summary	58
Chapter-4: A Cluster-based solution for Imbalance Data	59
4.1 Introduction	59
4.2 Preliminaries and basic definitions	60
4.2.1 Imbalanced Data	60
4.2.2 Between - class imbalance & within- class Imbalance	62
4.2.3 Imbalance Ratio	63
4.2.4 Degree of imbalance distribution	63
4.2.5 False positive and false negative	63
4.2.6 Clustering	64
4.2.7 SVM Classifier	64
4.3 Methods of handling Imbalanced Data	65
4.4 Experimental investigations	67
4.4.1 Datasets	67
4.4.2 Experiment Setting	67
4.5 Proposed Cluster Based Under-sampling	69
4.6 Methodology	70
4.7 Summary	84
Chapter-5: Feature Selection through Clustering to Classify High Dimensional Data	85
5.1 Introduction	85
5.2 Introduction about Data set	88

5.3	Preliminaries and basic definitions.....	89
5.3.1	SVM Classifier.....	89
5.3.2	RELIEF Feature Selection Approach.....	89
5.3.3	Info-Gain Feature Selection Approach.....	90
5.5	Methodology.....	90
5.6	Performance analysis of the Generated Model.....	96
5.7	Summary.....	104
Chapter-6: A Hybrid Model to Enhance the Performance of a Classifier.....		106
6.1	Introduction.....	106
6.2	Classification.....	107
6.2.1	SVM.....	107
6.3	Clustering.....	108
6.3.1	K-Means.....	108
6.4	Models for Data mining Tasks.....	109
6.4.1	WEKA.....	111
6.5	Datasets Description.....	112
6.6	Methodology.....	113
6.6.1	Data Preparation.....	113
6.6.2	Clusters Building.....	113
6.2.3	Building the classification Models:.....	113
6.3	Algorithm.....	115
6.4	Performance analysis of the generated model.....	117
6.5	Summary.....	123
Chapter-7: Conclusion and Future Work		124
7.1	Conclusion.....	124
7.2	Future Enhancements.....	125
Publications		127
References.....		128