Table of Contents

Chapter 1 .... **1**

1 Introduction 1

1.1 SMS Spam classification – An overview 1

1.2 Motivation 2

1.3 Aim of the work 3

1.4 Objectives of the Project 3

1.5 Organization of report 3

Chapter 2 .... **4**

2 Background 4

2.1 SMS 4

2.2 Reasons for growth in SMS spam 4

2.3 Machine Learning 5

2.3.1 Supervised Learning 6

2.3.2 Unsupervised Learning 6

2.3.3 Semi – Supervised Learning 7

2.4 Ensemble Learning 8

2.4.1 Supervised Ensemble Learning 8

2.4.1 Unsupervised Ensemble Learning 9

2.5 Feature Selection 9

2.5.1 Filter Methods 9

2.5.2 Wrapper Methods 9

2.6 Validation 10

2.6.1 Precision and Recall 10

2.6.2 K Fold Cross Validation 11

2.6.3 ROC Curve 12

2.7 Discussion 12

Chapter 3 13

Machine Learning Algorithms Used

3.1 Definition 13

3.2 Decision Tree 13

3.2.1 Definition 13

3.2.2 C4.5 Algorithm 14

3.2.3 Overfitting and Tree Pruning 15

3.3 Naïve Base 15

3.3.1 Introduction 15

3.3.2 Basic Principle 16

3.4 K-NN Algorithm 17

3.4.1 Description of Algorithm 17

3.4.2 Issue with K-NN 17

3.5 Supported Vector Machine 18

3.5.1 Introduction 18

3.5.2 Definition 18

3.6 Discussion 19

Chapter 4

Existing Classification Mechanisms for SMS spam Classification

4.1 Existing Solution 20

4.2 Discussion 22

Chapter 5

Proposed Method

5.1 Proposed Framework 23

5.2 Feature Extraction 24

5.3 Data Normalization 25

5.4 Feature Selection 27

5.4.1 Filter Method 27

5.4.2 Wrapper Method 28

5.4.3 Using Fuzzy C means 29

5.4.4 Hybrid approach using FCM and Naïve Base 32

5.5 Ensemble Classifier 33

5.6 Discussion 34

Chapter 6

Performance Study 35

6.1 Simulation Environment 36

6.2 Result and Analysis 36

6.2.1 Result over Full data set 36

6.2.1.1 Comparison of Accuracy of different Classifiers 36

6.2.1.2 Model Evaluation based on ROC curve 37

6.2.2 Result over Reduced Data set 38

6.2.2.1 Feature Selection using Filter method 39

6.2.2.2 Feature selection using Wrapper method 41

6.2.2.3 Feature selection using FCM 42

6.2.2.4 Feature selection using FCM and Wrapper using Naïve base 45

6.3 Discussion 47

Chapter 7

Conclusion and Framework 48

7.1 Conclusion 48

7.2 Future Work 48

Chapter 8

References 49

Figures

Chapter 2 4

2.1 Precision Explanation 10
2.2 Recall Explanation 10

Chapter 5 23

5.1 Framework for spam filtering 23
5.2. Normalized dataset using min-max normalization 26
5.3 System architecture of selecting features using FCM 31

Chapter 6 35

6.1 General User interface of weka 3.6 36
6.2 ROC curve of Different algorithm used in classification of ham with their ensemble.....37
6.3 ROC curve of Different algorithm used in classification of Spam with their ensemble .38
6.4 Accuracy of Classifier after gradual removal of features of low rank 39
6.5 Accuracy of classifier after gradual removal of features of low rank.............................39
6.6 Accuracy of classifier after gradual removal of features of low rank for SVMAttEval..40
6.7 Summarizes the Comparative accuracy 41
6.8 Accuracy of a classifier gradually removal of low ranked for all three evaluator...........41
6.9 Accuracy of Classifier using Feature selection with FCM using Euclidian 42
6.10 Accuracy of classifier after feature selection using FCM with std. Euclidian...............43
6.11 Accuracy of Classifier after Feature selection using FCM using std. City block..........44
6.12 Accuracy of Classifier after Feature selection using FCM using City block 44
6.13 Accuracy of proposed classifier using FCM with different distance metrics................45
6.14 ROC curve for different classifiers used and proposed ensemble classifier for ham 46
6.15 ROC curves for different classifiers used and proposed ensemble classifier for spam .46

List of Tables

Table 5.1. List of Structural Features 25

Table 5.2. Ranking of Structural Features 28

Table 5.3. Accuracy after removal of features using filters 28

Table 5.4. Accuracy of proposed classifier after feature selected based on mentioned classifiers 29

Table 5.5. Accuracy of classifier after removal of features using fuzzy c-means 32

Table 5.6. Accuracy After using Hybrid approach of feature selection using different distance metrics .. 32

Table 5.7. Accuracy of different classifiers before feature selection 33

Table 5.8. Accuracy of different classifiers after feature selection using Naïve Bayes................................34

Table 5.9. Accuracy of different classifiers after feature selection using FCM and Naïve Bayes 34

Table 6.1. Accuracy of different classifiers on prepared dataset without any feature selection...............37

Table6.2 Accuracy of proposed classifier after feature selection with wrapper method using different classifiers 42

Table 6.3. Accuracy of classifier after gradual removal of features using fuzzy c-means with different distances 42