

A STUDY ON SPAM CLASSIFICATION USING MACHINE

LEARNING TECHNIQUES



सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

A Dissertation Submitted in Partial Fulfilment of the Requirements for the
Fellow Programme in Management

**Indian Institute of Management
Indore**

By

Shrawan Kumar Trivedi

**Submitted in
May 2015**

Thesis Advisory Committee:

Prof. Shubhamoy Dey (*Chair*)

Prof. Prabin K. Panigrahi (Member)
(*Member*)

Prof. Sanjog Ray
(*Member*)

Abstract

In today's automated world, to be competitive and sustainable in business, information sharing within the units of the organisation is necessary. Email is an essential and useful tool of rapid and cheap communication. It is now a popular medium to connect people with each other. On the other hand, Spam (also known as unsolicited bulk email) is a challenge for the organisations and the researcher because its size is increasing day by day. This rapid growth causes serious hitches, such as unnecessary filling of users' mailboxes, engulfing some important emails, consuming storage space and bandwidth as well as too much time consumption in sorting them.

Various email spam filtering systems have been implemented by different organisations and the Internet Service Providers (ISPs) but none is found perfect. The existing systems suffer from a number of problems (like installation cost with no guaranteed positive outcomes, high false positive rate, training and testing time etc.) which are directly and indirectly reflected in the cost incurred due to spam and inadequate spam filters.

Recently, content-based filters are gaining popularity in the spam classification domain. Various machine learning algorithms are used in this method to classify emails. This method works with the content of the documents that are extracted to distinguish unsolicited email and legitimate emails. A critical issue has been identified in this method i.e. misclassification of the legitimate emails. Sometimes legitimate emails carry valuable information for a user, and if such good emails misclassify as spam, it will create serious problems. To tackle this issue, False Positive rate is used that evaluate the rate of misclassified legitimate emails. For a robust and sensitive classifier, the False positive rate should be as low as possible (ideally zero).

The need of the researchers and developers is to construct such filtering system that can be acceptable for any organisation and the Internet service providers. Content-based spam filters are popular due to their customization capability where, pre-existing email content can be used to train such filters. The emailing pattern of each organisation differs hence they need to install such filtering system which can be trained by particular words/features of their own pre-existing corpus so that the performance accuracy can be maximised and the False Positive Rate can be minimised.

This research proposes a content-based email spam filtering system with a set of novel algorithms along with some existing algorithms at the levels of feature selection as well as learning. Various studies have been done in different chapters for achieving this goal where several comparative studies between Feature Selection Methods, Feature Subset Search Methods, and Machine Learning Classifiers have been performed, together with some novel approaches proposed. All the tests have been done on popular publically available corpuses. According to the analysis, Relief F is predicted to be the best Feature Selection Method amongst other methods compared whereas Greedy Stepwise Search was found to be the best

feature subset search method in terms of good performance accuracy and low False Positive Rate.

Further, among Machine Learning Classifiers, Support Vector Machine and Random Forest are identified to be excellent whereas Bayesian classifier proved its worth in this domain. Support Vector Machine together with Greedy Stepwise Feature Search was the best pair of the Machine Learning and Feature Subset Search Method. Results have shown that this pair was excellent in terms of good performance accuracy and low FP rate amongst other pairs.

For Improving the Performance of classifiers, various combining and ensemble-based techniques were also included in this research. For combining method, the best combination of classifiers (boosted Probabilistic classifiers and Support Vector Machine) has been used to construct a combined classifier with committee Selection Mechanisms.

A novel Ensemble based Approach with the Genetic Programming Classifier has been developed and tested together with the best classifiers. It has been identified that this Enhanced Genetic Programming (EGP) Classifier is promising from all the performance dimensions.

Finally, this research has achieved the objective that was proposed initially by developing two different filtering models that have a trade-off between Accuracy and Training Time. First model incorporates Enhanced Genetic Programming Classifier with Greedy Stepwise feature search method and proves to be robust, fast (in Testing Time), accurate, and sensitive with less false positive rate but with high Training Time compared to the other models. Second model, that has been developed with Boosted Bayesian classifier and Greedy Subset feature search method, is robust, fast (both in Training and Testing), and sensitive with less false positive rate but classification accuracy is less than the first model proposed. In addition, the proposed filtering models satisfactorily tackle the issue of customization so that organisations can install these models to train according to their need.

The Proposed Models will cater the need of Organisations and Internet Service Providers (ISPs) where after installation they can minimize the cost related to spam and spam filtering system.

This research has contributed in the literature of spam classification by the following way:

1. Observed and validated RF (Relief F) as a best feature selection method.
2. Observed and validated Greedy Stepwise search as a best feature subset search method.
3. Observed and validated NP (Normalised Polynomial) as a best kernel for SVM (Support Vector Machine).
4. Observed and validated AdaBoost as best boosting method.
5. Observed and validated SVM (Support Vector Machine) as a best machine learning classifier.

6. Observed and validated SVM and Greedy Stepwise search as a best combination of machine learning and feature subset search methods.
7. Developed and validated a novel combining classifier with committee selection method.
8. Developed and validated a novel EGP (Enhanced Genetic Programming) machine learning classifier.
9. Two new models have been proposed:
 - EGP with Greedy Stepwise search.
 - Bayesian with Greedy Stepwise search.

Keywords: Spam Classification, Machine Learning Classifiers, Feature Selection Methods, Feature Subset Search Methods, Ensemble of Classifiers, Enhanced Genetic Programming Classifier, Combining Classifiers, Performance Accuracy, F-Value, False Positive Rate, Training Time, Testing Time.

Acknowledgements

Though this dissertation belongs to me but a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my FPM experience has been one that I will cherish forever.

Foremost, I want to offer this endeavour to our God Almighty for the wisdom he bestowed upon me, the strength, peace of my mind and good health in order to finish this research.

My deepest gratitude is to my advisor, Prof. Shubhamoy Dey. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. Prof. Dey taught me how to question thoughts and express ideas. His patience and support helped me to overcome many crisis situations and finish this dissertation. I am thankful to him for encouraging the use of correct grammar and consistent notation in my writings and for carefully reading and commenting on countless revisions of this Thesis. I am grateful to him for holding me to a high research standard and enforcing strict validations for each research result, and thus teaching me how to do research. I hope that one day I would become as good as an advisor to my students as Prof. Dey has been to me.

My co-advisors, Prof. Prabin Kumar Panigrahi and Prof. Sanjog Ray, have been always there to listen and give advice. I am deeply grateful to both of them for the long discussions that helped me sort out the technical details of my work. Prof. Panigrahi's insightful comments and constructive criticisms at different stages of my research were thought-provoking and they helped me to focus my ideas.

I am grateful to Prof. Rajhans Mishra, Prof. Saini Das and Prof. Madhukar Dayal for their encouragement and practical advice. I am also thankful to them for reading my research, commenting on my views and helping me understand and enrich my ideas.

I would like to mention some other names like FPM Chair Prof. Ranjeet Nambudari, Prof. Ganesh Kumar, Prof. Prashant Salwan, Prof. Abha Chatterjee for the many valuable discussions, suggestions and guidance about my writing and research that helped me to improve myself.

I am also grateful to the following former or current staff of FPM at IIM Indore, for their various forms of support during my doctoral study—Mr. Mukesh, Mr. Sandeep Das and Ms. Monika Mandloi.

I would like to acknowledge my friends, Asit Acharya, Anuj Sharma, Sriranga Vishnu, Ankit Sharma, Hemant Shrivastava, Arun Giri, Avik Sinha, Kapil Kaushik, Baljeet Alok, Sripad Sudhir, and many more, have helped me to stay sane through these difficult years. Their support and care helped me to overcome setbacks and stay focused on my studies. I greatly value their friendship and I deeply appreciate their belief in me.

Most importantly, none of this would have been possible without the love and patience of my wife Ankita. My loving wife, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength in my difficult situations. I would like to express my heart-felt gratitude to my wife for being with me at my toughest time.

Table of Contents

Chapter 1. Introduction	4
1.1. Spam Emails.....	5
1.2. Evolution of Spam Emails.....	7
1.2.1. The Early Year (Manual Spamming).....	7
1.2.2. The Second Phase (Machines for Spamming)	7
1.2.3. Third Phase (Machine against Machine)	8
1.3. Types of Spam	8
1.3.1. Advertisement Spam.....	9
1.3.2. Financial Spam.....	14
1.3.3. Phishing.....	16
1.3.4. Image Spam	17
1.4. Spam Consequences.....	17
1.5. Spam Filters.....	19
1.6. Spam Filtering Methods	19
1.6.1. Non Technical Solutions....	20
1.6.1.1. Recipient Revolts	20
1.6.1.2. Customer Revolts.....	21
1.6.1.3. Vigilante Attack.....	21
1.6.1.4. Hiding the Addresses.....	21
1.6.1.5. Legitimate Contacts and Limiting Trial Accounts.....	22
1.6.2. Technical Solutions.....	22
1.6.2.1. Domain Filters.....	23
1.6.2.2. Black Listing.....	23
1.6.2.1. White Listing.....	23
1.6.2.2. Rule Based Methods.....	24
1.7. Motivation.....	25
1.8. Research Objective.....	26
Chapter 2. Literature Survey	29
2.1. Manual Inspection	29
2.2. System Approaches	30
2.2.1. Grey Listing.....	31
2.2.2. White Listing.....	31
2.2.3. Black Listing.....	31
2.2.4. Collaborative approach.....	31
2.2.5. Challenge Responce.....	32
2.3. Content based Methods	33
2.3.1. Ad-hoc Rule Based Approach	33
2.3.2. Bayesian Filtering	34
2.4. Machine Learning Research.....	34
2.4.1. Bayesian Classifiers.....	34
2.4.2. Perceptron.....	37
2.4.3. Support Vector Machine.....	39
2.4.4. Nearest Neighbors.....	43
2.4.5. Decision Tree.....	45
2.4.5.1. Alternating Decision Tree.....	46
2.4.5.2. Decision Stump.....	46

2.4.5.3.	Reduced Error Pruning.....	47
2.4.6.	Random Forest.....	47
2.5.	Approaches to Create a Better Classifier	50
2.5.1.	Boosting Algorithms.....	51
2.5.1.1.	Bagging.....	51
2.5.1.2.	Boosting with Re-Sampling.....	52
2.5.1.1.	Adaptive Boosting.....	53
2.5.2.	Combining Classifiers.....	54
2.5.3.	Evolutionary Algorithms.....	56
2.5.3.1.	Genetic Algorithm based classifiers.....	57
2.5.3.2.	Genetic Programming based classifiers.....	57
2.5.4.	Research on Different Part of Spam	62
2.6.	Current Anti-Spam Systems.....	62
2.6.1.	Government Initiatives for Anti-Spam	63
2.6.2.	Industry Oriented Anti-Spam Associations	64
Chapter 3.	Experimental Design	67
3.1.	Structure of Spam Filter	67
3.2.	Different Types of Attacks on Email	69
3.2.1.	Tokenisation.....	69
3.2.2.	Obfuscation	69
3.2.3.	Weak Statistical	69
3.2.4.	Strong Statistical	69
3.3.	Corpora.....	69
3.3.1.	Most Complex Enron Corpus	70
3.3.2.	SpamAssassin	70
3.3.3.	LingSpam.....	71
3.3.4.	Training with Enron (5, 6) and Testing with Enron (All Version)	71
3.4.	Pre-processing	72
3.4.1.	Feature Extraction.....	72
3.4.2.	Dimensionality Reduction	73
3.4.3.	Feature Selection Process	73
3.4.3.1.	Document Frequency.....	73
3.4.3.2.	Information Gain.....	74
3.4.3.3.	Gain Ratio.....	74
3.4.3.4.	Chi-Square.....	75
3.4.3.5.	Relief F.....	75
3.4.3.6.	One Rule.....	76
3.4.4.	Feature Subset Search	76
3.4.4.1.	Genetic Search.....	77
3.4.4.2.	Greedy Stepwise search.....	78
3.4.4.3.	Best First Search.....	78
3.4.5.4.	Rank Search.....	79
3.4.5.	Re-parameterisation	79
3.4.5.1.	Latent Semantic Indexing.....	79
3.5.	Feature Representation.....	80
3.6.	Evaluation Parameters.....	81
3.7.	Discussion	83
Chapter 4.	Feature Selection and Subset Search Methods.....	84
4.1.	Section 1: Evaluation of Best Feature Selection Technique	84
4.1.1.	Aim of This Study.....	84

4.1.2.	Corpora For This Study	85
4.1.3.	Feature Selection Techniques	85
4.1.4.	Classifier for This Study	85
4.1.5.	System Design	85
4.1.6.	Evaluation Metrics.....	85
4.1.7.	Results and Analysis	85
4.2.	Section 2: Feature Subset Search	96
4.2.1.	Aim of This Study.....	96
4.2.2.	Corpora For This Study	96
4.2.3.	Feature Subset Search Techniques.....	96
4.2.4.	Classifiers for This Study.....	96
4.2.5.	System Design	97
4.2.6.	Evaluation Metrics.....	97
4.2.7.	Results and Analysis	97
4.3.	Discussion	100
Chapter 5. Machine Learning Classifier		102
5.1.	Evaluation of Best Machine Learning.....	102
5.1.1.	Aim of This Study.....	102
5.1.2.	Corpora For This Study	102
5.1.3.	Feature Subset Search Techniques.....	102
5.1.4.	Classifiers for This Study.....	103
5.1.5.	System Design	103
5.1.6.	Evaluation Metrics.....	103
5.1.7.	Results and Analysis	103
5.2.	Discussion	106
Chapter 6. Machine Learning with Excellent Features.....		108
6.1.	Best Combination of Machine Learning and Features Subset Selection	108
6.1.1.	Aim of This Study.....	108
6.1.2.	Corpora For This Study	108
6.1.3.	Feature Subset Search Techniques.....	108
6.1.4.	Classifiers for This Study.....	109
6.1.5.	System Design	109
6.1.6.	Evaluation Metrics.....	109
6.1.7.	Results and Analysis	109
6.2.	Discussion	115
Chapter 7. Combining and Ensemble Based Classifiers.....		116
7.1.	Section 1: Combining Classifiers with Committee Selection	116
7.1.1.	Aim of This Study.....	117
7.1.2.	Corpora For This Study	117
7.1.3.	Feature Subset Search Techniques.....	117
7.1.4.	Classifiers for This Study.....	117
7.1.5.	System Design	117
7.1.6.	Evaluation Metrics.....	118
7.1.7.	Results and Analysis	118
7.2.	Discussion	131
7.3.	Section 2: Enhanced Genetic Programming Classifier	133
7.3.1.	Aim of This Study.....	134
7.3.2.	Corpora For This Study	134
7.3.3.	Feature Subset Search Techniques.....	135
7.3.4.	Classifiers for This Study.....	135

7.3.5.	System Design	135
7.3.6.	Evaluation Metrics.....	135
7.3.7.	Results and Analysis	136
7.4.	Discussion	140
Chapter 8.	Training, and Testing Time	141
8.1.	Evaluation of Robust Email Filtering Models.....	141
8.1.1.	Aim of This Study.....	141
8.1.2.	Corpora For This Study	142
8.1.3.	Feature Subset Search Techniques.....	142
8.1.4.	Classifiers for This Study.....	142
8.1.5.	System Design	142
8.1.6.	Evaluation Metrics.....	142
8.1.7.	Results and Analysis	143
3.4.	Discussion	151
Chapter 9.	Conclusion, Business Implication and Limitation	152
9.1.	Business Implications.....	153
9.2.	Contribution of this Research.....	154
9.3.	Future Work	155
9.4.	Limitations	155
References.....		157
Appendix A:	Features selected by Re-Parametrization Methods	170
Appendix B:	Features selected by Feature Selection Method (Relief F).....	172
Appendix C:	Features selected by Feature Subset Search Method (Greedy Stepwise).....	174

List of Tables

Chapter 1: Introduction.....	4
Table 1. Statistics of Spam.....	18
Table 2. Categories of Spam	18
Chapter 2: Literature Survey.....	29
Table 3. Research work Related to Manual Inspection Methods	30
Table 4. Research work Related to System Response based Approaches	33
Table 5. Research work Related to Probabilistic Classifiers	36
Table 6. Research work Related to Perceptron based Algorithms	38
Table 7. Kernel Functions.....	41
Table 8. Research work Related to Support Vector Machine	42
Table 9. Research work Related to Neural Network based Classifiers.....	44
Table 10. Research work Related to Tree based Algorithms.....	49
Table 11. Research work Related to Stacking and Combining Classifiers.....	55
Table 12. Research work Related to Evolutionary Algorithms	59
Table 13. Research work Related to different part of an Email filebased Algorithms	61
Table 14. Current Spam filtering systems.....	65
Chapter 3. Experimental Design	67
Table 15. Corpora Descriptions	72
Table 16. Term Document Binary Representation	81
Table 17. Performance Metrics	82
Chapter 4. Feature Selection and Subset Search Methods.....	84
Section 1.Evaluation of best Feature Selection Methods.....	84
Table 18. Percentage Accuracy for Enron corpus.....	86
Table 19. Percentage F-Value for Enron corpus.....	87
Table 20. Percentage Accuracy for SpamAssassin corpus	88
Table 21. Percentage F-Value for SpamAssassin corpus	89
Table 22. Percentage Accuracy for LingSpam corpus.....	90
Table 23. Percentage F-Value for LingSpam corpus	91
Table 24. Percentage FP-Rate for Enron corpus.....	92
Table 25. Percentage FP for SpamAssassin corpus	92
Table 26. Percentage FP Rate for LingSpam corpus	92
Section 2.Evaluation of best Feature Subset Search Methods.....	96
Table 27. Percentage Accuracy and F-Value for all corpuses	98
Table 28. Percentage False Positive Rate for all corpuses.....	99
Chapter 5. Machine Learning Classifiers.....	102
Table 29. Percentage Accuracy and F-Value for all corpuses	104
Table 30. Percentage False Positive Rate for all corpuses.....	105
Chapter 6. Machine Learning with Excellent Features.....	108
Table 31. Percentage Accuracy and F-Value for Enron corpus.....	109
Table 32. Percentage Accuracy and F-Value for SpamAssassin corpus	110
Table 33. Percentage Accuracy and FP-Rate for LingSpam corpus.....	110
Table 34. Percentage FP Rate for all corpuses.....	114
Chapter 7. Combining and Ensemble based Classifiers	108
Section 2. Combining Classifiers with committee selection	116
Study 1 st . Boosting of the Probabilistic Classifiers.....	118
Table 35. Accuracy and F-Value of Probabilistic Classifiers	119
Table 36. FP Rate (Ham and all emails) of Probabilistic Classifiers	119
Study 2 nd . Kernel Selection for Support Vector Machine (SVM)	118

Table 37. Accuracy and F-Value of SVM with different Kernels	124
Table 38. FP Rate (Ham and all emails) of SVM with different Kernels	125
Study 3 rd . Combining classifiers with committee selection.....	128
Table 39. Accuracy and F-Value of Combining Classifier	128
Table 40. FP Rate (Ham and all emails) of Combining Classifier	128
Section 2: Enhanced Genetic Programming Classifier	133
Table 41. Parameters' Value for Novel EGP Classifier.....	135
Table 42. Accuracy and F-Value for all corpuses.....	136
Table 43. False Positive Rate for all corpuses	139
Chapter 8 : Training and Testing Time.....	141
Table 44. Percentage Accuracy for Machine Learning Classifiers.....	143
Table 45. Percentage F-Value for Machine Learning Classifiers	144
Table 46. False Positive Rate for Machine Learning Classifiers	146
Table 47. Training Time for Machine Learning Classifiers	147
Table 48. Testing Time for Machine Learning Classifiers	149

List of Figures

Chapter 1: Introduction	4
Fig 1. Different Kinds of Spam Sets	6
Fig 2. Marketing Spam.....	9
Fig 3. Online Pharmacy Spam.....	10
Fig 4. Stock Encouraging Spam	10
Fig 5. Pornographic or (Sex-) Dating Spam.....	11
Fig 6. Pirate Software Spam	11
Fig 7. Online Casino Spam.....	12
Fig 8. Fake Degree Spam	12
Fig 9. Mule Job Spam.....	13
Fig 10. Cause Promotions Spam	13
Fig 11. Fraud based Spam	14
Fig 12. Lottery Spam.....	15
Fig 13. Virus Spam.....	15
Fig 14. Phishing Spam.....	16
Fig 15. Image Spam.....	17
Chapter 2: Literature Survey.....	29
Fig 16. Structure of an Email message.....	60
Chapter 3. Experimental Design	67
Fig 17. Structure of an Email message and Spms Filter	68
Chapter 4. Feature Selection and Subset Search Methods	84
Section 1. Evaluation of best Feature Selection Methods.....	84
Fig 18. Percentage Accuracy for Enron corpus.....	87
Fig 19. Percentage F-Value for Enron corpus	87
Fig 20. Percentage Accuracy for SpamAssassin corpus	89
Fig 21. Percentage F-Value for SpamAssassin corpus	89
Fig 22. Percentage Accuracy for LingSpam corpus.....	91
Fig 23. Percentage F-Value for LingSpam corpus.....	91
Fig 24. Percentage FP-Rate for Enron corpus	93
Fig 25. Percentage FP for SpamAssassin corpus	93
Fig 26. Percentage FP Rate for LingSpam corpus	93
Section 2. Evaluation of best Feature Subset Search Methods	96
Fig 27. Percentage Accuracy for all corpuses	98
Fig 28. Percentage F-Value for all corpuses	98
Fig 29. Percentage False Positive Rate for all corpuses	99
Chapter 5. Machine Learning Classifiers.....	102
Fig 30. Percentage Accuracy and F-Value for all corpuses	105
Fig 31. Percentage False Positive Rate for all corpuses.....	105
Chapter 6. Machine Learning with Excellent Features.....	108
Fig 32. Percentage Accuracy and F-Value for Enron corpus.....	110
Fig 33. Percentage Accuracy and F-Value for SpamAssassin corpus	112
Fig 34. Percentage Accuracy and FP-Rate for LingSpam corpus	112
Fig 35. Percentage FP Rate for all corpuses.....	114
Chapter 7. Combining and Ensemble based Classifiers	108
Section 2. Combining Classifiers with committee selection	116
Study 1 st . Boosting of the Probabilistic Classifiers.....	118
Fig 36. Accuracy and F-Value of Probabilistic Classifiers (Enron)	120
Fig 37. FP Rate of Probabilistic Classifiers (Enron)	120

Fig 38.	Accuracy and F-Value of Probabilistic Classifiers (SpamAssassin)	121
Fig 39.	FP Rate of Probabilistic Classifiers (SpamAssassin).....	121
Fig 40.	Accuracy and F-Value of Probabilistic Classifiers (LingSpam)	123
Fig 41.	FP Rate of Probabilistic Classifiers (LingSpam)	123
Study 2 nd .	Kernel Selection for Support Vector Machine (SVM)	118
Fig 42.	Accuracy and F-Value of SVM with different Kernels (Enron)	124
Fig 43.	FP Rate of SVM with different Kernels (Enron)	125
Fig 44.	Accuracy and F-Value of SVM with different Kernels (SpamAssassin)	126
Fig 45.	FP Rate of SVM with different Kernels (SpamAssassin).....	126
Fig 46.	Accuracy and F-Value of SVM with different Kernels (LingSpam)	127
Fig 47.	FP Rate of SVM with different Kernels (LingSpam)	127
Study 3 rd .	Combining classifiers with committee selection	128
Fig 48.	Accuracy and F-Value of Combining Classifier (Enron)	129
Fig 49.	FP Rate (Ham and all emails) of Combining Classifier (Enron)	129
Fig 50.	Accuracy and F-Value of Combining Classifier (SpamAssassin).....	130
Fig 51.	FP Rate (Ham and all emails) of Combining Classifier (SpamAssassin).....	130
Fig 52.	Accuracy and F-Value of Combining Classifier (LingSpam)	131
Fig 53.	FP Rate (Ham and all emails) of Combining Classifier (LingSpam)	131
Section 2:	Enhanced Genetic Programming Classifier	133
Fig 54.	Percentage Accuracy for all corpuses	137
Fig 55.	Percentage F-Value for all corpuses	137
Fig 56.	False Positive Rate for all corpuses	139
Chapter 8 :	Training and Testing Time	141
Fig 57.	Percentage Accuracy for Machine Learning Classifiers	144
Fig 58.	Percentage F-Value for Machine Learning Classifiers	144
Fig 59.	False Positive Rate for Machine Learning Classifiers	147
Fig 60.	Training Time for Machine Learning Classifiers	147
Fig 61.	Testing Time for Machine Learning Classifiers	149

List of Abbreviations

Abbreviations	Meaning
AdaBoost	Adaptive Boosting
AD Tree	Alternative Decision Tree
ANN	Artificial Neural Network
BB	Boosted Bayesian
BF	Best First
BNB	Boosted Naive Bayes
CLX	Commercial Internet Exchange Association
CS	Chi-Square
DF	Document Frequency
DS	Decision Stump
EGP	Enhanced Genetic Programming
FP	False Positive
GA	Genetic Algorithms
GP	Genetic Programming
GR	Gain Ratio
GS	Genetic Search
GSS	Greedy Stepwise Search
IDS	Intrusion Detection Systems
IG	Information Gain
IPS	Intrusion Protection Systems
IS	Information Systems
ISP	Internet Service Provider
KNN	k-Nearest Neighbors
M3AAWG	Messaging Malware Mobile Anti Abuse Working Group
MAAWG	Messaging Anti Abuse Working Group
ML	Machine Learning
MTA	Mail Transfer Agent
MUA	Mail User Agent
NA	Not Applicable
NB	Naive Bayes
NN	Nearest Neighbors
NP	Normalized Polynomial Kernel
OR	One Rule
PK	Polynomial Kernel
PUK	Pearson VII function based Universal Kernel
RBF	Radial Basis Function Kernel
REP	Reduced Error Pruning
RF	Random Forest
RF	Relief F
RS	Rank Search
SVM	Support Vector Machine