A STUDY ON SPAM CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

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

List of Tables

List of Figures

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