How to cite this paper:

Mi ZhiWei, Manmeet Mahinderjit Singh, & Zarul Fitri Zaaba. (2017). Email spam detection: A method of metaclassifiers stacking in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 750-757). Sintok: School of Computing.

EMAIL SPAM DETECTION: A METHOD OF METACLASSIFIERS STACKING

Mi ZhiWei, Manmeet Mahinderjit Singh, Zarul Fitri Zaaba
School of Computer Sciences, University Sains Malaysia (USM) , 11800 Penang, Malaysia
mizhiwei@outlook.com, manmeet@usm.my, zarulfitri@usm.my

ABSTRACT. Nowadays, email is one of the fastest ways to conduct communications through sending out information and attachments from one to another. Individuals and organizations are all benefit the convenience from email usage, but at the same time they may also suffer the unexpected user experience of receiving spam email all the time. Spammers flood the email servers and send out mass quantity of unsolicited email to the end users. From a business perspective, email users have to spend time on deleting received spam email which definitely leads to the productivity decrease and cause potential loss for organizations. Thus, how to detect the email spam effectively and efficiently with high accuracy becomes a significant study. In this study, data mining will be utilized to process machine learning by using different classifiers for training and testing and filters for data preprocessing and feature selection. It aims to seek out the optimal hybrid model with higher accuracy or base on other metric's evaluation. The experiment results show accuracy improvement in email spam detection by using hybrid techniques compared to the single classifiers used in this research. The optimal hybrid model provides 93.00% of accuracy and 7.80% false positive rate for email spam detection.

Keywords: Email Spam Detection, MetaClassifiers, Stacking, False-Positive Rate, Accuracy

INTRODUCTION

Spam is "unsolicited bulk email" (Hidalgo, 2002), which "information crafted to be delivered to a large number of recipients, in spite of their wishes." Cormack (2007) defined spam with advertising content or fraud content are sent out in the way of mass mailing. Generally, mostly recognized form of spam is email spam. But, the forms of spam could be various according to the different media spam activities utilized, such email spam, SMS spam, Usenet newsgroup spam, social networking spam, Internet forum spam, file sharing spam and so on. All of these forms could be considered as spam. Therefore, to realize the impact of spam activities and to detect the email spam with advanced technology are becoming more and more necessary. Previously, there are many research study done by using data mining technique such as data mining via classification. Throughly much effort emphasise on single classifier. But however, spamming activities are changing the tactics to avoid the spam detection. Cisco security research recently stated that"a small number of spear-phishing messages purporting to originate from Apple Inc., claiming that the recipients had downloaded a popular game for mobile iOS devices. The email subject line included a randomly generated receipt number, another seemingly authentic touch, since legitimate emails would usually contain such a number''(Cisco annual security report). So, only using the simple classification may not have sufficient power to detect email spam since it requires more and more training data for machine learning to catch up the expanding step of spam. A solution through hybrid data mining could be research in depth especially in tackling email spamming issue. So far, hybrid technique has been explored in a number of different domains to meet the challenging task of analysis of large movement datasets (Huang & Chen, 2006; Tretyakov, 2004; Sivakami, 2015; Hemanth & Doreswamy, 2012; Vaishali & Ritesh, 2014)

Thus, in this research, we will focus in depth on technique for tackling email spam by using data mining technique. Questions such as: Whether the hybrid model provides better accuracy result comparing to any single classifier used for email spam detection will be observed through experimentation. Another important metric will be catered within this research is the false positive rate within the hybrid models. The objectives of this research are: i) To analyze single classifiers and hybrid models for email spam detection and evaluate the accuracy, performance and other evaluation metrics; ii) To present an analysis and benchmark hybrid modelsand generate the optimal combination of a single classifier with filter for email spam detection and iii) To provide analysis and benchmark hybrid models of two classifiers by stacking method and generate the optimal combination of two classifiers. The significance of this study is the select the best optimal hybrid approach and technique that can be adopted by system administrators within the organization in tackling the severity of impact of the email spamming attack. The outline of the paper includes state of art, research methodology, experiment results, discussions and findings, limitation of the study and conclusion.

STATE OF ART

In this section, a background of mobile computing, email security attacks and data mining detection techniques are brief account.

Mobile Computing & its Security Challenges

Mobile computing is defined as "an umbrella term used to describe technologies that enable people to access network services anyplace, anytime and anywhere" (Pattnaik & Mall, 2015). Mobile computing associated with several advantages, such as: locational flexibility, enhance productivity, and the emergence of cloud computing, etc. However, there also have challenges that mobile computing has to face. Several challenges of mobile computing are addressed by Forman and Zahorjan (1994) such as: low bandwidth, low power, and security risks..There are several ways of security attack on email. One of them is viruses. The most common way for viruses is to replicate and spread it through the email. Once the viruses enter into a mobile device, it may destroy data, operation system, and even the entire host system. Second one is phishing. It is the activity that phisher attempts to acquire email users' sensitive information such as username, password, credit card, bank and online payment information. It is the process to target at certain email users and by sending emails attached with a fake website in order to get those kinds of information. Those fake websites look like the normal real one and try to require users to enter personal information. Spam is another issue for email security. By doing the spam activities, spammers can flood the email system and affect system availability. Furthermore, the spam email may carry with viruses and Trojan horses. Once the email users failed to distinguish that spam email and download attachments from spam email. Then, the computers and systems have highly chance to get infection. Organizations will lose money and lead other potential cost when spam flood its email network and also waste their resource. Employees have to spend time and effort in order to distinguish spam or legitimate email if the email filtering system unable to detect spam with higher accuracy.In the past, the major email spam detection is listed based, such as setup blacklist, blackholes and so on. Nowadays, the spam activities become more and more intelligent. In order to distinguish and detect spam email, content based method is developing and becoming popular. For example, by using the data mining techniques.In this next section we will discuss regarding data mining technique to tackle email spamming attacks

Email Text Classification and Email Spam Detection

For the purpose of email spam detection by using data mining techniques, classification and clustering can be used. But in real fact, the filtering systems detect the email spam one by one in order to classify them and labelled as "legitimate" or "spam". Therefore, the classification is more suitable to facilitate email spam detection. The aim of text classification is to assign "natural language texts to one or more thematic category on the basis of their contents" (Zainal et.al, 2015). For the purpose of text classification, there are several machine learning algorithms can be used such as Naïve Bayes, Decision Tree and Support Vector Machine (SVM), etc. Since the experimental environment may differ, it cannot say that which specific classifier is the best. Different classifiers have its own characteristic and advantages. But, in general, the Decision Tree, SVM and Naïve Bayes are the most welcomed classifiers when process the machine learning for text classification, such as email spam detection.

Classification for Spam Detection Using Single Classifier.

When utilizing the data mining techniques, it is necessary to understand that each classifier may have advantage facing certain context and different classification work. Since not all the attributes of data are useful and meaningful for machine learning and data mining processing. If there are many irrelevant and redundant attributes in dataset can result in overfitting, much more time and memory consumption. So, successful feature selection of original attributes becomes more important. Currently, many data mining tools/application provide feature selection which can contribute to the model accuracy and decrease the model training time. In Rathi and Pareek'studey (2013), two experiments are processed. One without features selection and another with selecting features. By using the evaluator of "Best-First", the accuracy performance of different classifiers have been improved. The high dimensional data makes training and testing of the classification task difficult, so feature selection plays an important role for the classification task.

Classification via Hybrid Technique for Spam Detection.

Previously, the research study done based on the different classifiers to conduct a comparison of these classifiers' performance. Some classifiers have very good accuracy performance. But however, spamming activities are changing the tactics to avoid the spam detection. So, only using the simple classification may not have sufficient power to detect email spam since it requires more and more training data for machine learning to catch up the expanding step of spam. A hybrid technique has been explored in a number of different domains to meet the challenging task of analysis of large movement datasets (Huang & Chen, 2006; Tretyakov, 2004; Sivakami, 2015; Hemanth & Doreswamy, 2012; Vaishali & Ritesh, 2014). Hybrid techniques are able to perform well by providing better accuracy within different domains compared with single classifier. Therefore, it is necessary to conduct experiments on several hybrid models to identify whether it is applicable for email spam detection or not

Table 1. Existing Research with Hybrid Techniques.

Existing Research	Hybrid Techniques	Outcome

Credit scoring with a data mining approach based on support vector machines (Huang & Chen, 2006)	"SVM + Grid" "SVM + Grid + F-score," "SVM + GA."	A hybrid of SVM + GA can obtain good classification performance
A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set (Tretyakov, 2004)	"SVM using Polynomial kernel function" "SVM using RBF kernel function" Multilayer Perceptron Radial Basis Function Network(RBFN)	SVM+ Polynomial perform the best accuracy
Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model (Sivakami, 2015)	Decision Tree-SVM	SVM has higher prediction accuracy than IBL and SMO.
Hybrid Data Mining Technique for Knowledge Discovery from Engineering Materials Data Sets (Hemanth & Doreswamy, 2012)	Naive Bayesian classifier with pearson correlation coefficient method	
Approaches of Opinion Mining and Performance Analysis: A Survey (Vaishali & Ritesh, 2014)	NB + GA	Hybrid technique provide much better accuracy than Genetic Algorithm and NB.

Refer to the summary from Table 1, hybrid techniques are able to perform well by providing better accuracy within different domains compared with single classifier. Therefore, it is necessary to conduct experiments on several hybrid models to identify whether it is applicable for email spam detection or not.

RESEARCH METHODOLOGY

In order to achieve the objectives of the research, the suitable dataset is required. For this research study, "spambase" dataset which collected by UCI Machine Learning Repository is adopted. It consists of 58 attributes where 57 continuous attribute and 1 nominal class label attribute and the total number of instances is 4601. Waikato Environment for Knowledge Analysis (WEKA) tool is adopted. For this research, based on the literature review, the Naïve Bayes, Support Vector Machine and Decision Tree are the most welcomed algorithms for text classification. Therefore, according to the classifier availability of WEKA, the Naïve Bayes, and Sequential Minimal Optimization from Support Vector Machine algorithms and J48 from Decision Trees algorithms are selected as classifiers for this research paper.For the experiment part of hybrid of classifiers with filters, the "Ranker" is choosen as the search algorithm for the feature selection. By doing so, the evaluators for the attributes of datasets are limited to "Information Gain", "Gain Ratio" and "Chi Square" in WEKA.

EXPERIMENTATION RESULTS

In order to evaluate the performance of classifiers with filter and feather selection, "Ranker" is used as the only one search algorithm for attributes to rank all attributes according to the attributes evaluation and process the attribute selection. The default setting of "numToSelect" in Weka for "Ranker" is -1 which means that WEKA will select all attributes for the next step of data classification task even the attribute evaluation has been done. Therefore, there is a need to set how many attributes should be selected manually. For this research, the data set consists 57 attributes, so that from top 10 until top 50 attributes will be selected and run separately with each classifier and filters. The accurate result of "Naïve Bayes", "SMO" and “J48” are 79.29%, 90.44% and 92.98% respectively. Meanwhile, the false positive rate of “Naïve Bayes”, “SMO” and “J48” are 15.2%, 12.1% and 7.8% respectively. The classifier “J48” performs the best result among these three single classifiers which has 92.98% accuracy and 7.8% false positive rate. Meanwhile, the classifier “Naïve Bayes” provides the worst result which accuracy is 79.29% and false positive rate is 15.2%.

The best result for future “Chi Square” with classifier is a hybrid model of “J48” as classifier, “Chi Square” with “Ranker” as filter and 50 attributes selected, which accuracy is 93.00% and 7.80% false positive rate. The worst accuracy and false positive rate are 79.57% and 28.20% when combine “Naïve Bayes” with “Chi Square” and select 50 attributes and 10 attributes correspondingly as the hybrid model to detect email spam. The best result for filter “Gain Ratio” with classifier is a hybrid model of “J48” as classifier, “Gain Ratio” with “Ranker” as filter and 30 attributes selected, which accuracy is 92.78% and 8.00% false positive rate. The worst accuracy is 71.29% when combine “Naïve Bayes” with “Gain Ratio” and select 20 attributes as the hybrid model to detect email spam. The worst false positive rate is 23.90% when combine “SMO” with “Gain Ratio” and select 10 attributes.

The best result for filter “Information Gain” with classifier is a hybrid model of “J48” as classifier, “Information Gain” with “Ranker” as filter and 50 attributes selected, which accuracy is 92.83% and 7.80% false positive rate. The worst accuracy and false positive rate are 79.57% and 27.30% when combine “Naïve Bayes” with “Information Gain” and select 50 attributes and 10 attributes correspondingly as the hybrid model to detect email spam. For the part of hybrid model of two classifiers by stacking method detects the spam email by a combination of two classifiers, it means that two classifiers selected are known as the base classifiers to be used. Each base classifier will be trained by giving the training dataset in Weka, and these two base classifier will provide the different “estimate” result. After that, the stacking method will build and train a “metaLearner” between these two base classifiers and their previous “estimate” results become as input. And then, this “metaLearner” will generate the final classification result based on “metaClassifier” selected within these two base classifiers.

Table 2: Results Summary of Two Classifiers by Stacking Method.

Base Classifiers	“MetaClassifier”	Accuracy	False Positive Rate
SMO+Naïve Bayes	SMO	90.44%	12.1%
SMO+Naïve Bayes	Naïve Bayes	81.87%	14.4%
SMO + J48	SMO	92.98%	7.8%
SMO + J48	J48	92.89%	8.1%
J48+Naïve Bayes	J48	93.22%	8.3%
J48+Naïve Bayes	Naïve Bayes	93.18%	8.7%

Table 2 shows the results summary of two classifiers by using stacking method. The accuracy and false positive rate differ when select different base classifier as the “metaClassifier” in Weka. The best accuracy result is 93.22% by using the hybrid model of “J48” with “Naïve Bayes” and “J48” performs as “metaClassifier”. The best false positive rate is 7.8% by using the hybrid model of “SMO” with “J48” and “SMO” performs as “metaClassifier”. The certain hybrid models provide better accuracy and false positive rate result by comparing to single classifier’s performance.

DISCUSSION & FINDINGS

The performance of single Naïve Bayes classifier has the worst result among those three classifiers which only has 79.29% accuracy. This may due to the drawback of Naïve Bayes classifier itself and the limitation of dataset. Naïve Bayes assumes that all features of dataset are independent which means each attribute is independency among dataset. But in fact, when solving the certain practical issues, they do have dependency among features, so that the drawback of Naïve Bayes leads the loss of accuracy performance. Also, the data set utilized has limited instances which may not produce the best possible results of Naïve Bayes classifier. However, by utilizing the hybrid models of Naïve Bayes with filters, the classification performance has slight improvement. When hybrid of "Naïve Bayes" with "Information Gain" and 30 attributes selected, the accuracy result improves 10 percent by comparing to single Naïve Bayes performance. Table 3 shows the summary of the best hybrid models of each filters. J48 is the classifier provide the best performance results of each kind hybrid of single classifier with different filters.

Table 3. Summary of the Best Hybrid Models of Each Filters.

Hybird Models			Accuracy Improvement	False Positive Rate
J48	IG	50	-0.15%	7.80%
J48	GR	30	-0.20%	8.00%
J48	Chi	50	0.02%	7.80%

For the best performance one of a hybrid model of J48 with filter, 93.00% accuracy and 7.8% false positive rate performed by "J48" with "Chi Square" and 50 attributes selected. Comparing to the single J48 classifier's accuracy performance, the hybrid model one leads a 0.02 percent improvement for email spam detection accuracy. For this research, the dataset has 4601 instances which means hybrid model one will detect extra 92 units of email as spam or legitimate email correctly. Thus, if the dataset and amount of instances becomes huge which able to reflect the practical fact in business organization environment, then the hybrid model will definitely show its advantage to handle spam email detection.

J48 performs well when it is used as a hybrid model of classifier with filters is because the method decision tree algorithms build the "Tree" which is based on the probability of occurrence by testing the each attributes of instances. For this research, the numbers of attributes of dataset is relative small which J48 is able to compute the probability more accurately. All the same time, the filters reduce certain numbers of attributes which lead J48 able to build the "Tree" better and facilitate the email classification process.For the part of hybrid of single classifier with various filters, the best hybrid models for each filters have been summarized in Table 3. The hybrid model of "J48" with "Chi Square" and 50 attributes selected has the highest accuracy and false positive rate. Therefore, it is the best hybrid model among the part of hybrid of single classifier with various filters. Why "Chi Square" facilitate "J48" classifier to generate much better results comparing to "Information Gain" and "Gain Ratio"? The theory of "Chi Square" is used in statistic to test the independence of two events. "Chi Square" as a filter of hybrid classification model will test whether the occurrence of a specific attribute and the occurrence of a specific class are independent or not. Because usually a small part of the attributes of the dataset are independent of class. Thus, "Chi Square" facilitates to evaluate all the attributes and select the meaningful features from classification tasks.If based on the literature reviews, the feature selection will facilitate the classification performance since the irrelevant and redundant attributes are eliminated.

The accuracy increase and hit a peak point with certain number attribute selected. After that, the accuracy will keep decline following the growth of number of attribute selected. The result of this hybrid model improves 0.24 percent of accuracy by comparing to the best result of single classifier, but this hybrid model has 0.5 percent decrease in false positive rate (shown in Table 4).

Table 4. Summary of the Best Hybrid Model of Two Classifiers byStacking.

Base Classifiers	“metaClassifier"	Accuracy	False Positive Rate
SMO+J48	SMO	92.98%	7.80%
J48+Naïve Bayes	J48	93.22%	8.30%

The best result of false positive rate is achieved by SMO with J48 and SMO performs as “metaClassifier”, which is equal to the best false positive result of single classifier and the hybrid model of classifier with filter. By using the stacking method to hybrid two classifiers, SMO with Naïve Bayes has the worst accuracy result which is 81.87% when Naïve Bayes performs as “megaClassifier” which is still due to the drawbacks of Naïve Bayes itself. Therefore, for these two hybrid models, which one is considered as the best hybrid of two classifiers by stacking method? “J48” with “Naïve Bayes” provides the best accuracy result, but its false positive rate is slightly higher. False positive means a legitimate email detected as a spam email by the filtering system. A high false positive rate will cause cost and potential opportunity loss for the business organizations. Thus, the best hybrid model among the part of two classifiers by stacking method is “SMO” with “J48” and “SMO” performs as “metaClassifier”, which accuracy is 92.98% and false positive rate is 7.80%.

CONCLUSION & FUTURE WORK

By conducting this research, it generates a best hybrid model for email spam detection which leads accuracy improvement. From a business perspective, SMEs are able to apply this kind of hybrid model to build their email filtering system to detect the spam and reduce the impact of spam activities. For this research, it aims to evaluate and seek out a hybrid model technique of data mining to handle the problem of email spam. As a result, “J48” with “Chi Square” and 50 attributes selected is considered as the optimal hybrid model for this research which provide 93.00% accuracy and 7.80% false positive rate to detect email spam.The research shows hybrid models bring the performance improvement compared to single classifier. Techniques such as clustering or prediction should be future observed as well.

REFERENCES

Cisco Annual Security Report. Retrieved from https://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2015_ASR.pdf

Cormack, G. V. (2007). Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4), 335-455.

Forman, G. H., & Zahorjan, J. (1994). *The challenges of mobile computing. Computer*, 27(4), 38-47.

Hemanth, K. S., & Doreswamy. (2012). Hybrid Data Mining Technique for Knowledge Discovery from *Engineering Materials Data Sets*.

Hidalgo, J. M. G. (2002, March). Evaluating cost-sensitive unsolicited bulk email categorization. In *Proceedings of the 2002 ACM symposium on Applied computing* (pp. 615-620). ACM.

Huang, C. L., & Chen, M. C. (2006). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*

Pattnaik, P. K., & Mall, R. (2012). Fundamentals of Mobile Computing. PHI Learning Pvt. Ltd..

Rathi, M., & Pareek, V. (2013). Spam Mail Detection through Data Mining-A Comparative Performance Analysis. *International Journal of Modern Education and Computer Science*, 5(12), 31.

Sivakami, K. (2015). Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, Vol.1.

Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. In *Data Mining Problem-oriented Seminar, MTAT* (Vol. 3, No. 177, pp. 60-79).

Vaishali, M. & Ritesh K. S. (2014). Approaches of Opinion Mining and Performance Analysis: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4.

Zainal, K., Sulaiman, N. F., & Jali, M. Z. (2015). An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka. *International Journal of Computer Science and Information Security*, 13(3), 66.