ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Effective Spam Filtering using Random Forest Machine Learning Algorithm

Manish Kumar

Researcher, Department of Computer Science, Banaras Hindu University, Institute of Science, UP, India

ABSTRACT: Now-a-days e-mail isbecoming a fast and economical facility to exchange information. However, unwanted or junk e-mail also known as spam became a foremost problem on the today's Internet and isresponsible for financial damage to companies, irritating individual users, wasting the network resources andthe most important have become an cumulative problem for informationSecurity. To solve these problems the users of e-mail should have automated tool that can filter the spam e-mails automatically. In this study, the experiments were conducted for spam e-mailsfiltering task on the dataset obtained from UCI Machine Learning Repository separately using tenmachine learning algorithms with ten-fold cross validation. The result obtained shows that classifier Random Forest is outperforming with AUC, accuracy and MCC value up to 0.987**,** 0.955 and0.906 respectively.

KEYWORDS: Spam e-mails, Machine Learning, UCI Machine Learning Repository, AUC, Random Forest.

I. INTRODUCTION

Electronic mail, also known as e-mail, is a most popular, fastest and cheapest way of exchanging digital messages using Internet andbecoming anintegral part of everyday life for millions of people.The improvement and explosion of information and network technologies lead theorganizations and individuals progressively moretrust on emails to communicate and share information and knowledge. Even though they may delight and enjoy this effective, useful medium, individuals and organizations also agonise from spam e-mails, which have increased theatrically in number in recent times. These Spam, or unwanted commercial or bulk email, is not requested by recipients but sent to the inboxof recipient[1, 2, 3]. The spam emails not only consume users' time and energy to recognise and eliminate the undesired messages, but also become originof many frustrating problems for instance taking up restricted mailbox space, overwhelming important individual emails, and wasting network bandwidth [3]. Also, spam emails even can be destructive for children which contain pornographic materials [3,4]. Since there is no cost for sending emails, resultantlyenormously increasing volume of spam emails hasoccurred, andby using address harvesting tools spammers can obtain email addresses easily. Jupiter Research [5] approximate that 4.9 trillion spam emails were sent worldwide in 2003. Onemodern report guesses that spam emails have increased from approximately 10% of overall mail volume in 1998 to as much as 80% today [6,7]. Another investigation by Fallows [8] results that 52% of email users specifythat spam has made them less credulous of email communication services, and 25% express that the volume of spam has reduced their interest in email usage. To reduce the costs and increase the credibility of the user effective spam filtering is required, which automatically discriminates spam from legitimate emails, can be essential to both individuals and organizations.

II. RELATED WORK

Fetterly, Manasse, and Najork [9,10] analysed the content properties of spam pages by using statistical methods however**Ntoulas, Najork, Manasse, and Fetterly** [11] used machine learning approaches for identifying spam content. **Erdelyi, Garzó, and Benczúr** [12] study offered a widespreadinvestigation of how several content features and machine-learning models canadd the quality of a web spam detection algorithm. Consequently, in addition to feature selection, effective classifiers were built using boosting, bagging and oversampling, [13; 14]. Newly, **Prieto, Álvarez, López-García, and Cacheda** [15] offered a system called SAAD, in which web content is used to identify web spam.**Page, Brin, Motwani, and Winograd** [16] introduced solution for link spam using PageRank and HITS methods and the solutions introduced by **Kleinberg** [17] are reflected the first best solutions to fightagainst the web spam. Since then, different proposals have been specifically focused on link spam by introducing many alternatives to

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

detect it [18]. **Wu and Davison** [19] and **Chellapilla and Maykov** [20] evaluated web redirection spam. With reference to click spam, an excitingsolution for its stoppagewas proposed by **Radlinski** [21].Furthermore, the work of **Immorlica, Jain, Mahdian, and Talwar** [22] considered the problem of click fraud for online advertising platform whereas **Prieto** et al. [23] introduced an incentive based ranking model. **Geng, Wang, Li, Xu, and Jin** [24] presented the first study which used both content and link-based features to identify web spam pages. **Svore, Wu, Burges, and Raman** [25] and **Abernethy, Chapelle, and Castillo** [26] studied were concentrated on link and content-based features forbuildingperfect classifiers using SVM. **Rungsawang, Taweesiriwate, and Manskasemsak** [27] used ant colony algorithm to classify web spam and compared its result with SVM and decision trees. **Dai, Davison, and Qi** [28] used SVM and Logistic Regression for classification task of spams.**Becchetti, Castillo, Donato, Leonardi, and Baeza Yates** [29] combined link and content-based features using C4.5 to identify web spam. **Silva, Alimeida, and Yamakami** [30] investigated with numerousclassifiersincluding decision tree, SVN, KNN, LogitBoost, Bagging and AdaBoost in their analysis. **Araujo and Martínez-Romo** [31] introduced an effective spam discovery system founded on a classifier that associations link-based features with language-model characteristics. **Karimpour, Noroozi, and Alizadeh** [32] suggested a method constructed on the Expectation–Maximization algorithm with Naïve Bayes classifier to decide the labeling problem

III. MATERIAL AND METHODS

A. Dataset

For study, dataset has been downloaded from the UCI Machine Learning Repository named "Spambase" , having number of instances 4601 and number of features 58 (57 continuous and 1 nominal class label).

B. Features used

Word frequency: 48 continuous real features are selected in which frequency of few spam indicative WORDS (order, mail, receive, remove, credit and many more) is calculated by using the following formula

A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

> Percentage of words in the e-mail that match WORD=100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail.

Character frequency: 6 continuous real features are selected in which frequency of few spam indicative CHARACTERS (;, (, [, !, $ and #) is calculated by using the following formula

> Percentage of characters in the e-mail that match CHARACTER=100 * (number of times the WORD appears in the e-mail) / total number of characters in e-mail.

Average length of capital letters: The average length of uninterrupted sequences of capital letters is calculated and used as feature in spam filtration
Longest length of sequence of capital letters: Thelongest length of uninterrupted sequence of capital letters is calculated and used as feature in spam filtration.
Total length of Capital letters: Total number of capital letters in the e-mail are also used as feature.
Goal: Whether the e-mail was considered spam (1) or not (0).

C. Classifying protocol: Random Forest

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Random forests [33] are a combination of tree predictors so that all trees depend on the values of a random vector sampled autonomously and with the similar distribution for all trees in the forest. The random forests algorithm for prediction or classification task can be explained as follows:
1. Using original samples data draw n tree bootstrap
2. For every of the bootstrap samples, produce an unpruned classification tree, by following modification: at each node, instead of choosing the best split among all predictors, arbitrarily sample m try of the predictors and select the best split among those variables.
3. Predict new data by aggregating the predictions of the ntree trees using majority votes for classification.
An estimation of the error rate can be found, based on the training data, by the following steps:
1. At every bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of bag", or OOB, data) by considering the tree developed with the bootstrap sample.
2. Cumulate the OOB predictions. (On the average, every data point would be out-of-bag around 36% of the times, so cumulate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

IV. RESULTS AND DISCUSSION

For testing our proposed method the experiments were conducted for filtration task of spam by separately applying ten machine learning algorithms namely: Random Forest (RF), Average One-Dependence Estimators (AODE), Fisher's linear Discriminate Function (FLDA), Logistic Model Trees (LMT), LOGISTIC, Radial Basis Function Classifier (RBFC), Rotation Forest with J48 base Classifier (ROF+J48), Rotation Forest with LMT as base classifier (ROF+LMT), Simple Logistic(SLG) and Sequential Minimal Optimization(SMO) using Weka 3.7.12 [34]. The classification performances of the classifiers were analysed with respect to the standard performance parameters, namely: Accuracy, Specificity, Sensitivity, Precision, Receiver Operating Characteristic (ROC) Area [35], Matthew's Correlation Coefficient (MCC) besides time taken for training (learning). The formula for calculating these parameters are given below:

$$Sensitivity = \frac{tp}{tp + fn} * 100 \tag{1}$$

$$Specificity = \frac{tn}{tn + fp} * 100 \tag{2}$$

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{3}$$

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

$$MCC = \frac{(tp * tn) - (fp * fn)}{\sqrt{(tp + fn) * (tn + fp) * (tp + fp) * (tn + fn)}} \tag{5}$$

where

tp is the number of true positives,
tn is the number of true negatives,
fp is the number of false positives and
fn is the number of false negatives.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

The table 1 shows the values of Sensitivity, Specificity, Accuracy, Precision, MCC, AUC performance metrics besides their training time for all the twelve classifiers separately for our chosen dataset.

Classifiers	Sensitivity	Specificity	Accuracy	Precision	MCC	AUC	Training Time (in sec)
RF	0.972	0.929	0.955	0.955	0.906	0.987	004.13
AODE	0.956	0.901	0.934	0.934	0.862	0.980	000.20
FLDA	0.941	0.850	0.905	0.905	0.800	0.951	000.37
LMT	0.952	0.916	0.937	0.937	0.869	0.966	041.46
LOGISTIC	0.949	0.886	0.924	0.924	0.841	0.971	001.81
RBFC	0.917	0.809	0.874	0.874	0.735	0.935	003.70
ROF+J48	0.968	0.923	0.950	0.950	0.896	0.984	015.30
ROF+LMT	0.959	0.918	0.943	0.943	0.881	0.985	563.27
SLG	0.952	0.886	0.926	0.926	0.844	0.973	010.22
SMO	0.952	0.831	0.904	0.905	0.798	0.891	000.66

Table 1: Performance of ten classifiers for spam filtration task

The sensitivity indicates the ability of the classifier to identify positive instances correctly, the specificity indicates the ability of the classifier to identify negative instances correctly and accuracy indicates the percentage of correct classification of both positive class as well as negative class instances. The Random Forest performs better than other classifiers with sensitivity, specificity and accuracy values 0.972, 0.929and 0.955respectively.

The Mathews Correlation Coefficient (MCC) is another important parameter to evaluate the performance of the binary class classifiers. A coefficient of +1 represents a perfect classification, 0 an average random classification and −1 an inverse classification. It can be observed from the table 1 that, that classifier having high value of accuracy performance parameter for a particular family also have high MCC. In our experiment the MCC value we achieved is 0.906for Random Forest.

The area under ROC curve (AUC) is an important statistical property to compare the overall relative performance of the classifiers. AUC can take values from 0 to 1. The value 0 for the worst case, 0.5 for random ranking and 1 indicates the best classification as the classifier has ranked all positive examples above all negative example. The figure 1 shows that AUC value of Random Forest classifier is greater than other classifier for our considered dataset equals to 0.987.

V. CONCLUSION AND FUTURE WORK

We have compared the performance of ten classifiers (including SVM, which was reported as the better performing classifier by the previous studies) for the filtering of spam task. The experimental results of our proposed method have demonstrated that RF has produced superior performance in terms of classification accuracy, AUC and MCC respectively for our considered dataset. It was also observed that few classifiers have yielded poor classification accuracy as compared to RF like SMO and RBFC. This problem will be investigated in our future study.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

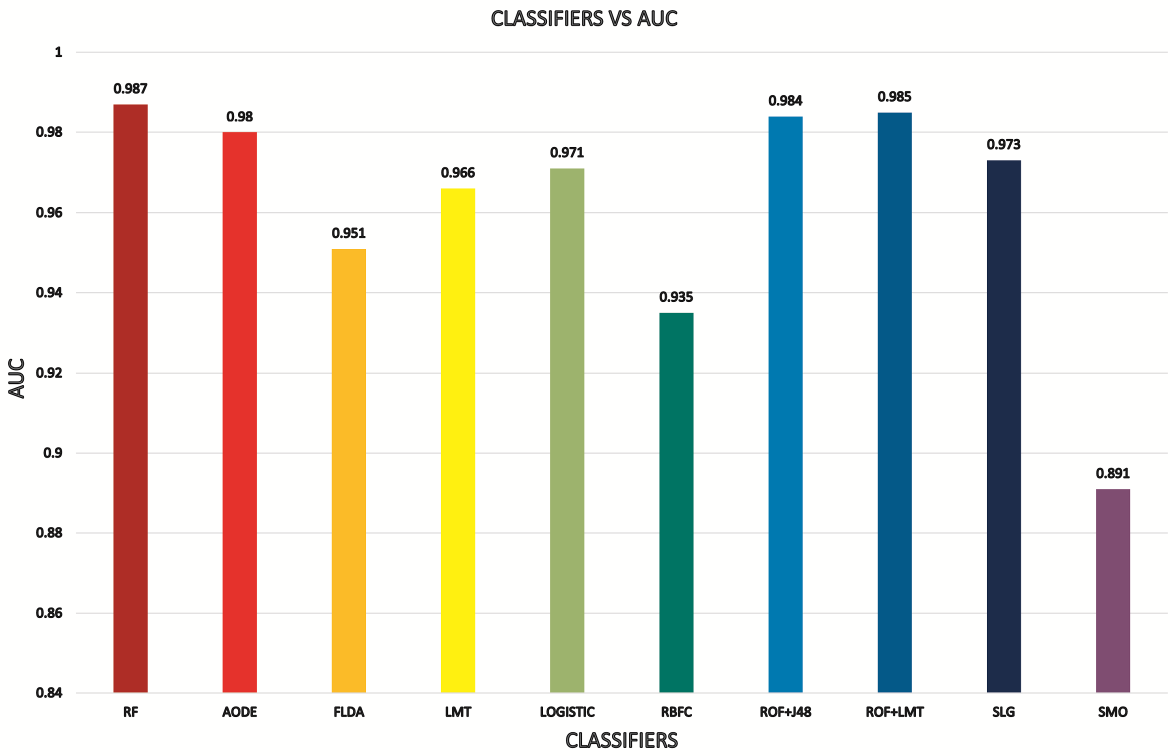

Fig 1: AUC of selected classifiers for spam filtration task

REFERENCES

[1]. P.O. Boykin, V.P. Roychowdhury, Leveraging social networks to fight spam, IEEE Computer 38 (4) (2005) 61–68.
[2]. B. Whitworth, E. Whitworth, Spam and the social-technical gap, IEEE Computer 37 (10) (2004) 38–45.
[3]. Zhang, J. Zhu, T. Yao, An evaluation of statistical spamfiltering techniques, ACM Transactions on Asian Language Information Processing 3 (4) (2004) 243–269.
[4]. V. Zorkadis, M. Panayotou, D.A. Karras, Improved spam e-mail filtering based on committee machines and information theoretic feature extraction, Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, 2005, pp. 179–184.
[5]. C. Taylor, Spam's big bang, Time (June 16, 2003) 51.
[6]. Messaging Anti-AbuseWorking Group, MAAWGEmail Metrics Program, First Quarter 2006 Report, June 2006, available at www.maawg.org/about/FINAL_1Q2006_Metrics_Report.pdf.
[7]. J. Goodman, G.V. Cormack, D. Heckerman, Spam and the ongoing battle for the inbox, Communications of the ACM 50 (2) (2007) 24–33.
[8]. D. Fallows, Spam: How It Is Hurting E-Mail and Degrading Life on the Internet, Technical Report (Pew Internet and American Life Project), October 2003, available at http://www.pewinternet. org/reports/toc.asp?Report=102.
[9]. Fetterly, D., Manasse, M., & Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In Proceedings of the seventh international workshop on the web and databases (WebDB'04) (pp. 1–6). Paris, France.
[10]. Fetterly, D., Manasse, M., & Najork, M. (2005). Detecting phrase-level duplication on the World Wide Web. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (pp. 170–177).
[11]. Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. In Proceedings of the 15th international conference on World Wide Web (WWW'06) (pp. 83–92).
[12]. Erdélyi, M., Garzó, A., & Benczúr, A. A. (2011). Web spam classification: a few features worth more. In Proceedings of the 2011 joint WICOW/AIRWeb workshop on web quality (WebQuality'11) (pp. 27–34). New York, USA.
[13]. Geng, G. G., Jin, X.-B., Zhang, X.-C., & Zhang, D. X. (2013). Evaluating web content quality via multi-scale features. International Journal of Computing Research Repository (CoRR). ArXiv: 1304.6181v1 [cs.IR].
[14]. Nikulin, V. (2010). Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In Proceedings of the ECML/PKDD 2010 discovery challenge.
[15]. Prieto, V. M., Álvarez, M., López-García, R., & Cacheda, F. (2012). Analysis and detection of web spam by means of web content. In IRFC. Lecture notes in computer science (Vol. 7356, pp. 43–57). Springer.
[16]. Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. In Proceedings of the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

seventh international World Wide Web conference (pp. 161–172). Brisbane, Australia.

[17]. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604–632.

[18]. Spirin, N., & Han, J. (2011). Survey on web spam detection: Principles and algorithms. ACM SIGKDD Explorations Newsletter, 13(2), 50–64.

[19]. Wu, B. & Davison, B. D. (2005b). Cloaking and redirection: A preliminary study. In Proceedings of the first international workshop on adversarial information retrieval on the web (AIRWeb 2005) (pp. 7–16).

[20]. Chellapilla, K., & Maykov, A. (2007). A taxonomy of javascript redirection spam. In Proceedings of the third international workshop on adversarial information retrieval on the web (AIRWeb 2007) (pp. 81–88) New York, USA.

[21]. Radlinski, F. (2007). Addressing malicious noise in clickthrough data. In In Proceedings of the third international workshop on Adversarial information retrieval on the web, AIRWeb'07. Banff, Canada.

[22]. Immorlica, N., Jain, K., Mahdian, M., & Talwar, K. (2005). Click fraud resistant methods for learning click-through rates. In Proceedings of the first international workshop on internet and network economics (WINE) (pp. 34–45). Hong Kong, China.

[23]. Prieto, V. M., Álvarez, M., López-García, R., & Cacheda, F. (2012). Analysis and detection of web spam by means of web content. In IRFC. Lecture notes in computer science (Vol. 7356, pp. 43–57). Springer.

[24]. Geng, G. G., Wang, C. H., Li, Q. D., Xu, L. & Jin, X. B. (2007). Boosting the performance of web spam detection with ensemble under-sampling classification. In Proceedings of IEEE fourth international conference on fuzzy systems and knowledge discovery (pp. 583–587).

[25]. Svore, K. M., Wu, Q., Burges, C. J. C., & Raman, A. (2007). Improving web spam classification using rank time features. In Proceedings of the third international workshop on adversarial information retrieval on the web (AIRWeb 2007) (pp. 9–16).

[26]. Abernethy, J., Chapelle, O., & Castillo, C. (2008). Webspam identification through content and hyperlinks. In Proceedings of the fourth international workshop on adversarial information retrieval on the web.

[27]. Rungsawang, A., Taweesiriwate, A., & Manskasemsak, B. (2012). Spam host detection using ant colony optimization. Lecture notes in electrical engineering (Vol. 107). Springer.

[28]. Dai, N., Davison, B. D., & Qi, X. (2009). Looking into the past to better classify web spam. In Proceedings of the fifth international workshop on adversarial information retrieval on the web (AIRWeb 2009) (pp. 1–8). New York, USA.

[29]. Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2008). Web spam detection: Link-based and content-based techniques. In Proceedings of the European integrated project dynamically evolving, large scale information systems (pp. 99–113). Heinz-Nixdorf-Institute.

[30]. Silva, R. M., Alimeida, T. A., & Yamakami, A. (2013). Machine learning methods for spamdexing detection. International Journal of Information Security Science, 2(3), 86–107.

[31]. Araujo, L., & Martínez-Romo, J. (2010). Web spam detection: New classification features based on qualified link analysis and language models. IEEE Transaction on Information Forensics and Security, 3(3), 581–590.

[32]. Karimpour, J., Noroozi, A. A., & Alizadeh, S. (2012). Web spam detection by learning from small labelled samples. International Journal of Computer Applications, 50(21), 1–5.

[33]. Breiman, L. (2001) Random forests. *Mach. Learning*, **45**, 5–32.

[34]. Witten H, Ian H. 2011. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Systems.

[35]. Tom Fawcett, (2003). ROC graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories.