ISSN 2278-3091

Volume 9, No.1.4, 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse3291.42020.pdf

https://doi.org/10.30534/ijatcse/2020/3291.42020

Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification

Doaa Mohammed Ablel-Rheem[1], Ashraf Osman Ibrahim[2,3], Shahreen Kasim[4],Abdulwahab Ali Almazroi [5], Mohd Arfian Ismail [6]

[1] Faculty of computer Science and Information Technology, Neelain University, Khartoum North, Sudan

[2] Faculty of computer Science and Information Technology, AlzaiemAlazhari University, Khartoum North, Sudan

[3] Arab Open University, Sudan Branch, Khartoum, Sudan

[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia

[5] Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Saudi Arabia

[6] Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Pahang, Malaysia

ABSTRACT

The data mining techniques produce good work in many domains. The spam emails are becoming a serious dilemma and an important matter to have different solutions, and enhanced methods and algorithms. Using Ensemble methods which are well-established classifiers. In this paper data mining techniques used to classify spam email using the UCI spam base dataset. The results achieved by the machine learning tools and techniques, and the Ensemble learning methods, after applying feature selection methods on the data set; which gave better result, and better classification accuracy. For the evaluation method used the cross-validation for testing and training option, and the confusion matrix to show the accuracy and the performance result of the chosen classifiers; which are Naïve Bayes, decision tree, ensemble boosting and ensemble hybrid boosting classifiers.

Key words : Email classification, Ensemble learning, Feature Selection Technique, Hybrid Ensemble

1. INTRODUCTION

Nowadays, Email is a part of millions of people's life. They use email for different purposes such as; business, study and for other reasons. It has changed the way man collaborates and works by being the most cheapest, popular and fastest means of communication [1]. It is a common amount for a user to receive hundreds of emails daily. Around 92% of these emails are spam [2]. They include advertisements for a variety of products and services, such as pharmaceuticals, electronics, software, jewelry, stocks, gambling, loans, pornography, phishing, and malware attempts [3]. The spam not only consumes the user's time by forcing them to identify the unwanted messages, but also wastes mailbox space, network bandwidth and time. Therefore, spam classification is becoming a bigger challenge to process for individuals and organizations [4] [23]. The word spam was used to describe unwanted, junk mails sent to an internet user's inbox. It is very convenient for spammers to send millions of email spam all over the world with no cost at all [5]. The filtering and detection techniques are the most commonly used methods; it identifies whether a message is spam or Non-SPAM based exclusively on the message content and some other characteristics of the message. Despite different approaches and techniques adopted to fight the scourge called spam, the internet today still witnesses huge amount of spam [5][6][8]. Even with the efforts that spent to reduce the SPAM emails, it is still considered to be a threat. Message Labs Intelligence reports that in 2010 the average global SPAM emails rate for the year was 89.1% an increase of 1.4% compared with 2009[7].

As the problem with which we are working is a classification problem, we not only need to have models that maximize the accuracy results of correct classified samples. We present a model that using feature selection method and ensemble classification. Feature selection to reduce the attribute of the email spam dataset and then ensemble classifier to detect and classify the results. In addition, the performance of the proposed hybrid model comparing the outcomes against some of the well-known classification techniques.

The paper is organized as follows: Section 2 provides some related studies. Section 3 describes the materials and methods. Section 4 gives the proposed method that used, followed by the experimental design and results obtained in Section 5. Finally, conclusions are drawn in Section 6.

2. RELATED STUDIES

There are some research works that apply data mining and machine learning methods and techniques in spam e-mail classification, the research [12] used four classifiers including Neural Network, SVM, Naïve Bayesian, and J48 were tested to filter spams from the dataset of emails. All the emails were classified as spam (1) or not (0).

The study [13] used the Word Stemming or Word Hashing Technique for improving the efficiency of the content based spam filter which is been employed in the SMTP server made a correct classification of the ham and spam emails. The study [14] proposes a new spam detection technique using the text clustering based on vector space model it computes disjoint clusters automatically using a spherical k-means algorithm for all spam/Not-SPAM mails and obtains centroid vectors of the clusters for extracting the cluster description. In other hand research [15] presents a new improved model that combines negative selection algorithm (NSA) with particle swarm optimization (PSO) has been proposed and implemented.

Study [16] explores and identifies the use of different learning algorithms for classifying spam messages from e-mail. A comparative analysis among the algorithms has also been presented. Through a comprehensive analysis of various classifiers using different software tools viz. WEKA, Rapid Miner was implemented on a common dataset. The researchers in [17] evaluate the performance of Non Linear SVM based Classifiers with various kernel functions over Enron Dataset and in order to evaluate as many as possible attributes SVM has proved to be good classifier because of its sparse data format and acceptable Recall and Precision Value. Also SVM is regarded as an important example of “kernel methods”, one of the key areas in machine learning. Authors in [18] describe classification of emails by Random Forests (RF) Algorithm and that ensemble learning provides a more reliable mapping that can be obtained by combining the output of multiple classifiers.

Other popular learning algorithms that have been applied to spam email detection include machine learning algorithms, feature selection methods [23] [24] [28] and also Naïve Bayes, SVM [25] [26] KNN classification [27].

3. MATERIAL AND METHODS

In this section, the dataset attributes is explained. Next, a brief review of notions and approaches of Feature Selection, ensemble learning, and evaluation methods. Finally, the proposed model is exhaustively described.

3.1 Feature Selection

Feature Selection (FS) used to overcome the task of extracting high dimensional data into the smallest possible [21]. The attribute selection give us ranking of every attribute describe the training data set. It helps us to choose the best attribute that giving best information to help on training process. The reduction of attribute let the training process more faster, reduce the amount of memory size used (less iteration process) and minimizes the expected number of tests needed to classify a given group. The three popular attribute selection measures are Information Gain, Gain Ratio and Gini Index [10]. In this paper Information Gain used as feature selection method and it is minimizes the amount of information needed to classify the tuples resulting partitions and reflects the least randomness or “entropy.” in these partitions.

3.2 Ensemble learning

It’s a supervised learning technique; it is a way to make the learner you got better. The Basic idea for ensemble classification is combining multiple models by building different “experts” and let them vote. One of the advantages of it often improves predictive performance, and it is a powerful machine learning paradigm which has shown better results in many applications. But, usually produces output that is very hard to analyze, and there are approaches that aim to produce a single comprehensible structure [10] [22]. One of the most popular methods is Boosting it is an alternative approach uses voting/averaging but weights models according to performance. It encourage new model to become an “expert” for instances misclassified by earlier models. Boosting needs weights but can adapt learning algorithm or can apply boosting without weights by resample data with probability determined by weights. The ensemble method that used is an adaptive Boosting (AdaBoost).

3.3 Evaluation Methods

To evaluate the model this following methods are used. For the testing and training option used cross-validation method, and the confusion matrix to show the accuracy and the performance result for each classifier [10].

- Cross-Validation In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2…Dk, each of approximately equal size. Training and testing is performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets D2…Dk collectively serve as the training set to obtain a first model, which is tested on D1; the second iteration is trained on subsets D1, D3…Dk and tested on D2; and so on. Unlike the holdout and random subsampling methods, here each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data. In general, stratified 10-fold cross-validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance [10].

- Confusion matrix a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Confusion matrix as shown in Table 1 can be represented in form of false positive (FP), false negative (FN), true positive (TP) and true negative (TN). The total number of tuples is TP +TN +FP +FN, or P + N [10].

$$Accuray = \frac{TP + TN}{P + N} \qquad (1)$$

Where is P+N = Total number of features.

Table 1: Confusion Matrix [10]

Actual class \Predicted class	C_1	$\neg C_1$
C_1	True Positives(TP)	False Negatives(FN)
$\neg C_1$	False Positives(FP)	True Negatives(TN)

Where:

- ✓ True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- ✓ True negatives (TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- ✓ False positives (FP): These are the negative tuples that were incorrectly labeled as positive. Let FP be the number of false positives.
- ✓ False negatives (FN): These are the positive tuples that were mislabeled as negative. Let FN be the number of false negatives.

3.4 Performance measure

Performance measure of each classification model is evaluated using statistical measure; classification accuracy. This measures are defined using True positive (TP), True negative (TN), False positive (FP), False negative (FN) [10].

- ✓ Precision exactness – what % of tuples that the classifier labeled as positive are actually positive.

$$Pr\,ecision = \frac{TP}{TP + FP} \qquad (2)$$

- ✓ Recall completeness – what % of positive tuples did the classifier label as positive? Perfect score is 1.0

$$Re\,call = \frac{TP}{TP + FN} \qquad (3)$$

- ✓ *F* measure (*F_1* or *F*-score) harmonic mean of precision and recall.

$$F - measure = \frac{2 * P * R}{P + R} \qquad (4)$$

Where:
P→ is the precision
R→ is the recall

4. PROPOSED MODEL

The proposed hybrid ensemble model is created from the combination of feature selection method, ensemble learning technique to improve the classification accuracy of the email spam dataset. Implementation phases of the proposed hybrid model are presented in Figure 1. The full details of theses phases are thereafter discussed.

Step 1. Starting with used SPAM base dataset and prepared the data to be read.
Step 2. Applying the feature selection techniques to select most important attribute of the spam data.
Step3. Implementing ensemble learning method as a classification methods; which been used hybrid boosting and REP-tree classifier technique.
Step 4. Finally choose the training and testing option to build the model and classifying spam and non-spam email.

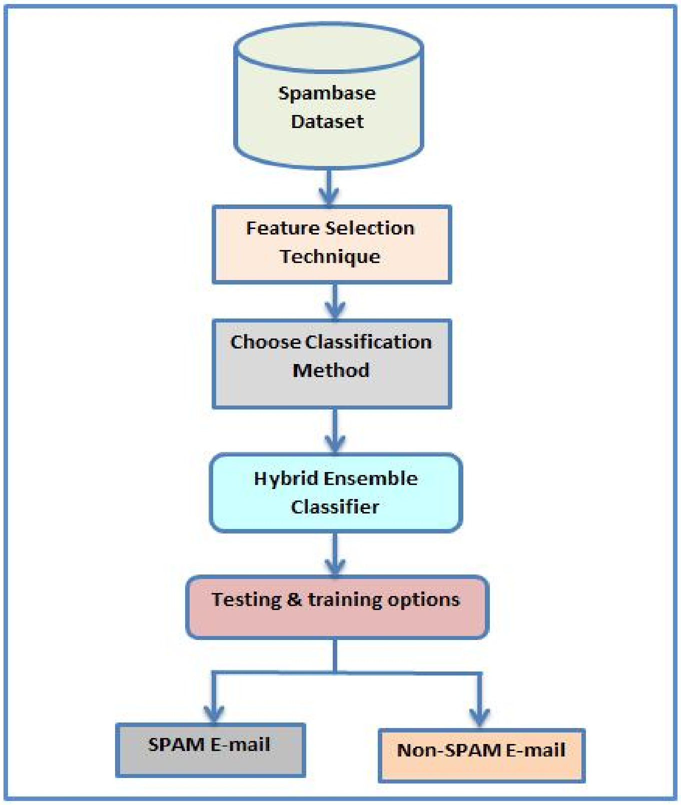
Spambase Dataset
Feature Selection Technique
Choose Classification Method
Hybrid Ensemble Classifier
Testing & training options
SPAM E-mail
Non-SPAM E-mail

Figure 1: The Proposed Ensemble Method

5. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental study of the hybrid ensemble model and feature selection algorithm. The results are compared to other classifiers. The details of these two data sets are given below.

5.1 Dataset Description

In this study we are using UCI Machine Learning Repository spambase dataset. This dataset has 4601instances and 57attributes and single output called class, this class represent the final output which is SPAM or Not-SPAM. The attributes are characterized to (Integer, and Real). The dataset has been donated by Gorge Farman since 1999. The "spam" concept is diverse for products/web sites to make money fast. The spam E-mail collection came from postmaster and individuals who had filed spam, but the Not-SPAM e-mails came from filed work and personal e-mails [20].

5.2 Experimental set-up

In order to show the process of the proposed method in some detail, first, the experiments for the email spam dataset which contain 57 attributes and 1 label class with total of 4601 instances and used confusion matrix and 10 fold cross-validation as well. We used Naïve Bayes classifier, decision tree and ensemble method for classification accuracy and to compare these methods with the proposed hybrid ensemble method. The proposed method used feature selection (FS) method to reduce the dataset attributes and the result was generated 40 attributes.

5.3 Result and discussion

This section show the results that achieved by the proposed method and other methods.

The proposed model built after; pre-processing the data set which is 57attributes and 1 label class, and applying feature selection (FS) method to reduce the dataset attributes to 40 and 1 label class.

Actual class \Predicted class	SPAM	Not-SPAM
SPAM	1725	88
Not-SPAM	865	1923

Table 2: Confusion Matrix Naïve Bayes Classifier

From table 2, The Naïve Bayes classifier correctly classified True Positive (TP) 1725 instances as spam email, and 1923 instances correctly classified as regular or not-spam email / True Positive (TN), and the 865 instances have been classified as spam but actually they are not False Negative (FN), and 88 instances has been classified as not-spam but actually they are spam email False Positive (FP).

Actual class \Predicted class	SPAM	Not-SPAM
SPAM	1646	167
Not-SPAM	156	2632

Table 3: Confusion Matrix Decision Tree Classifier

The Decision Tree classifier results shown in table 3, the correctly classified 1646 instances as spam email (TP), and 2632 instances correctly classified as regular or not-spam email (TN), and the 167 instances have been classified as spam but actually they are not (FP), and 156 instances has been classified as not-spam but actually they are spam email (FN).

Table 4: Confusion Matrix ensemble Classifier

Actual class \Predicted class	SPAM	Not-SPAM
SPAM	1566	247
Not-SPAM	210	2578

Table 4 shows the results achieved by ensemble classifier and we can see from the table the correctly classified 1566 instances as spam email (TP), and 2678 instances correctly classified as regular or not-spam email (TN), and the 210 instances have been classified as spam but actually they are not (FP), and 247 instances has been classified as not-spam but actually they are spam email (FN).

Table 5: Confusion Matrix hybrid ensemble Classifier

Actual class \Predicted class	SPAM	Not-SPAM
SPAM	1680	133
Not-SPAM	124	2664

The proposed hybrid ensemble model reported the results in table 5 with correctly classified 1680 instances as spam email (TP), and 2664 instances correctly classified as regular or not-spam email (TN), and the 124 instances have been classified as spam but actually they are not (FP), and 133 instances has been classified as not-spam but actually they are spam email (FN).

The correct classified Instances and incorrect classified Instances of the selected classifiers and the proposed method are listed in Table 6. The Naïve Bayes classifier achieved very low accuracy compared to the decision tree and ensemble classifier. While the proposed method obtained high classification accuracy compared to other classifiers used in this study. The feature selection method help the proposed method to achieve high classification accuracy and also enhanced classification results. In Table 6, the proposed method obtained the best results for correctly classified with high accuracy results with 94.41% and low incorrectly results with 5.58%.

Table 7: Result of the Classifiers

Classifiers	Correctly classified Instances (%)	Incorrectly classified Instances (%)
Naïve Bayes	79.28 %	20.71%
Decision Tree	92.97 %	7.02%
Ensemble	90.06 %	9.93 %
Hybrid Ensemble	94.41 %	5.58 %

The following Table 8 presents the evaluation measures for the chosen classifiers; which are the precision, recall and f-measure; presenting the weighted average for both classes (class 1 labeled for spam email and class 0 for the not-spam and). The following Figure 2 is graphic display for the evaluation measures of the four classifiers as shown in the Table 8.

Table 8: The Weighted Average of Evaluation Measures

Classifiers / Measures	Precision	Recall	F-Measure
Naïve Bayes	0.842	0.793	0.794
Decision Tree	0.930	0.930	0.930
Ensemble	0.900	0.901	0.900
Hybrid Ensemble	0.944	0.944	0.944

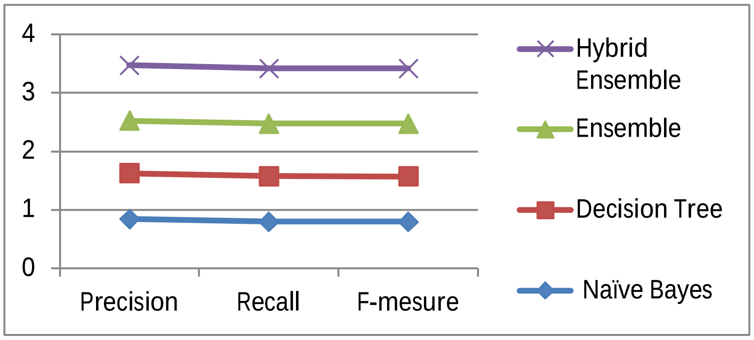

Figure 2: The Four Classifiers Evaluation Measures

The following Figure 3 presents the result of the selected classifiers accuracy which is the correctly classified instances; which shown in Table 7 previously.

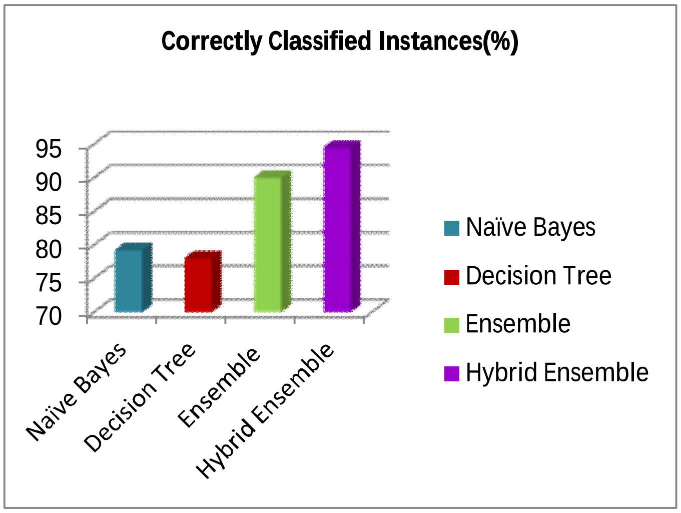

Figure 3: The Four Classifiers Accuracy

While Figure 4 presents the incorrectly classified instances gave by the selected classifiers which shown in Table 6 earlier.

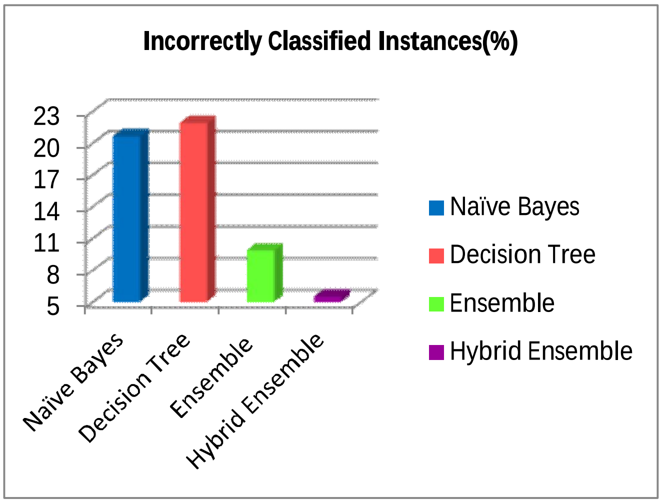

Figure 4: The Four Classifiers Incorrectly Classified Instances

6. CONCLUSION

Classification models based on hybrid machine learning methods have had a significant impact on the detection tasks. The aim of this study was emerging the accuracy of the classification models with take advantage of combining methods expected to emerge from the hybridization of the feature selection and ensemble method. The results achieved in this study by the Ensemble learning methods, after applying feature selection methods; gave better classification accuracy result. Bagging with Random subspace classifier gave much better accuracy result. Using the hybrid technique also improved the classifiers result. It also gave good values of precision and F-measure. Our further work will focus on more improvement to obtain highly accurate and interpretable classification accuracy. We try to use new technologies and more advanced hybrid swarm intelligence techniques.

ACKNOWLEDGEMENT

This research is supported by Geran Antarabangsa Vot W004 from Research Management Centre (RMC), Universiti Tun Hussein Onn Malaysia (UTHM).

REFERENCES

1. Whittaker, S., Bellotti, V. and Moody, P., 2005. Introduction to this special issue on revisiting and reinventing e-mail. Human-Computer Interaction, 20(1), pp.1-9.
2. DeBarr, D. and Wechsler, H., 2012. Spam detection using random boost. Pattern Recognition Letters, 33(10), pp.1237-1244.
3. Heron, S., 2009. Technologies for spam detection. Network Security, 2009(1), pp.11-15..
4. Zhang, Y., Wang, S., Phillips, P. and Ji, G., 2014. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 64, pp.22-31. https://doi.org/10.1016/j.knosys.2014.03.015
5. Zhang, L., Zhu, J. and Yao, T., 2004. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), pp.243-269.
6. Massey, B., Thomure, M., Budrevich, R. and Long, S., 2003, June. Learning Spam: Simple Techniques For Freely-Available Software. In USENIX Annual Technical Conference, FREENIX Track (pp. 63-76).
7. MessageLabs intelligence, Annual Security Report (2009) Retrieved 8th Feb, 2010.
8. Almeida, T.A., Almeida, J. and Yamakami, A., 2011. Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications, 1(3), pp.183-200.
9. Awad, W.A. and ELseuofi, S.M., 2011. Machine Learning methods for E-mail Classification. International Journal of Computer Applications, 16(1).
10. Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier. Third Edition, Jiawei Han University of Illinois at Urbana–Champaign.
11. Silva, C. and Ribeiro, B., 2007. RVM ensemble for text classification. International Journal of Computational Intelligence Research, 3(1), pp.31-35.
12. Youn, S. and McLeod, D., 2007. A comparative study for email classification. In Advances and innovations in systems, computing sciences and software engineering (pp. 387-391). Springer, Dordrecht.
13. Hamsapriya, T. and Renuka, M.D.K., 2010. Email classification for spam detection using word stemming. International journal of computer applications, 1(5), pp.45-47.
14. Sasaki, M. and Shinnou, H., 2005, November. Spam detection using text clustering. In null (pp. 316-319). IEEE. https://doi.org/10.1109/CW.2005.83
15. Idris, I. and Selamat, A., 2014. Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22, pp.11-27.
16. Scholar, M., 2010. Supervised learning approach for spam classification analysis using data mining tools. Organization, 2(08), pp.2760-2766.
17. Chhabra, P., Wadhvani, R. and Shukla, S., 2010. Spam filtering using support vector machine. Special Issue IJCCT, 1(2), p.3.
18. Bhagyashri U. Gaikwad, P. P. Halkarnikar Spam E-mail Detection by Random Forests Algorithm, 2013.
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.

20. Spambase.documentation at the UCI Machine Learning Repository,http://www.ics.uci.edu/~mlearn/MLRepository.html, May 01, 2018, 06:54:55 pm.
21. Mohamad, M. and Selamat, A., 2015, April. An evaluation on the efficiency of hybrid feature selection in spam email classification. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on (pp. 227-231). IEEE.
22. Hall, M.A., Practical Machine Learning Tools and Techniques. United State: Morgan Kauffman, 2011.
23. Shah, N.F. and Kumar, P., 2018. A Comparative Analysis of Various Spam Classifications. In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications(pp. 265-271). Springer, Singapore.
24. Bassiouni, M., Ali, M. and El-Dahshan, E.A., 2018. Ham and Spam E-Mails Classification Using Machine Learning Techniques. Journal of Applied Security Research, 13(3), pp.315-331.
25. Sah, U.K. and Parmar, N., 2017. An approach for Malicious Spam Detection In Email with comparison of different classifiers.
26. Kumaresan, T. and Palanisamy, C., 2017. E-mail spam classification using S-cuckoo search and support vector machine. International Journal of Bio-Inspired Computation, 9(3), pp.142-156.
27. Sharma, A. and Suryawanshi, A., 2016. A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. International Journal of Computer Applications, 136(6), pp.28-35. https://doi.org/10.5120/ijca2016908471
28. Kumar, R.K., Poonkuzhali, G. and Sudhakar, P., 2012, March. Comparative study on email spam classifier using data mining techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 14-16).