The Combination of Logistic Regression and Gradient Boost Tree for Email Spam Detection

Arfah Anggraini¹, Rifkie Primartha¹*, Adi Wijaya²

¹Faculty of Computer Science, Sriwijaya University, Indonesia
²Informatics Engineering Department, Universitas MH Thamrin, Jakarta, Indonesia
E-mail: arfah231197@gmail.com, rifkie@ilkom.unsri.ac.id, adiwjj@gmail.com
*corresponding author

Abstract. Convenience communicate in this era making trends throughout the world. Email is one of many tools that still used for communicating. In the activity of sending-receiving email, some irresponsible people often send spam messages to recipients only for profit. The process of spam filtering is still continue developed to reduce the number of spam emails. In this research, we propose Logistic Regression with Select by Weight and Gradient Boost Tree for developed spam filtering to make spam filters more advanced. The model has been built using Logistic Regression with Select by Weight and Gradient Boost Tree showing a good result. Accuracy generated from the mentioned models is 95.13%.

1. Introduction
Email is one of the many technology-based communication tools that are still used today. Email provides convenience for someone to connect to another. Various kinds of data that can be sent via email such as text, photos, videos and files. Sometimes we receive emails that we don't ask from people we don't know. This happens because the email can come from anywhere and at any time with the web that is easily accessible and fast in the market [1]. This unsolicited email is called spam.

Spam is a threat to recipients because spam could contain viruses and spyware that make the receiving system destroyed [2]. Spam could also contain an advertisement that can cause problems with the recipient's bandwidth [3]. In 2016, it was recorded that 61.66% of spam hampered the flow of world [4]. To overcome this, the role of spam filtering is needed to reduce the entry of an unwanted email for users.

Various kinds of software related to spam filtering have emerged as well as continuing to improve performance in order to work as well as possible. Previous research related to spam email detection has yielded very good results. The differences in the proposed method and the results of previous research will be shown in chapter IV. Result Experimental.

For increasing accuracy, we propose combining the Gradient Boost Tree (GBT) and Logistic Regression (LR) method with Select by Weight to detect spam emails. LR is implemented because it can overcome noise at GBT which can lead to overfitting.

Gradient Boost Tree Algorithm is an algorithm that can be used in the classification and Regression process [5]. The working method of the Gradient Boost Tree algorithm uses logic where the next
predictor will learn from previous predictors. The sequential addition of trees to GBT will correct the errors that occurred in the previous tree [6].

The Logistic Regression (LR) algorithm is a combination of techniques for showing and evaluating several factors emphasizing the relationship between innate variables and an assessment strategy for solving data sets where there is at least one autonomic factor that determines outcomes [7]. This algorithm is linear in predicting probability [8] and can be used for reducing noisy data [9]. The purpose of this model is to obtain a regression equation by predicting two or more groups of objects [10].

2. Related Works
Related Research on detection of spam email has been widely applied. With the same dataset from the UCI Machine Learning Repository, several research references were taken to compare the results that obtained from proposed method. These studies as in 2014 conducted by Zhou et al propose the method is cost-sensitive three-way (Bayesian, thresholds, probability) [11]. The process of splitting data in their research split the data into 80% for training data and 20% for testing data. The accuracy obtained from the proposed method is 89.88%.

Other studies related to the detection of spam email that has been done by Idris et al [12]. The method proposed in his research uses a combination of Negative Selection Algorithm (NSA) with Differential Evolution (DE). Differential Evolution (DE) is implemented because it is able to change the next detector at the stage of the Negative Selection Algorithm (NSA) [9]. The combination of Negative Selection Algorithm (NSA) with Differential Evolution (DE) produces an accuracy of 83.06%.

At the same year, 2014 [13] proposed the method to be used namely the Negative Selection Algorithm (NSA) with Particle Swarm Optimization (PSO). Particle Swarm Optimization (PSO) was used to produce a detector during the training of the Negative Selection Algorithm (NSA) algorithm. The results of the accuracy obtained from the proposed method are 91.22%.

Other studies in 2016 conducted by [14] used the same model as [13]. Other Particle Swarm Optimization (PSO) can correct random detectors of Negative Selection Algorithm (NSA). By using the same dataset get an accuracy of 83.20%.

Research conducted by [9] in 2016 gives a good result. They use a method that is a combination of Logistic Regression with FN Threshold (LRFNT) and Decision Tree (DT) to detect spam emails. Logistic Regression is used to reduce noisy data before being trained by the Decision Tree algorithm. Data is split into 70% for training data and 30% for testing data. By using the same dataset from the UCI Machine Learning Repository, the accuracy obtained is quite high 91.67%.

Other research uses the Fuzzy Granular approach by [15] with granular classification to form hyperbox to detect spam emails. Data splitting into 70% training data and 30% testing data. Accuracy obtained from the use of the Fuzzy Granular approach is 94.62%.

3. Materials and Proposed Method
In this study, we will propose the method used is a combination of Logistic Regression (LR) with Select by Weight and Gradient Boost Tree (GBT). The dataset used is from UCI Machine Learning Repository [16] with the dataset name is Spambase. These emails come from personal email and field work [17].

The spambase dataset has 4601 data. The data has 57 attributes with 56 regular attributes and 1 special attribute. Special attributes indicate that email is categorized as spam (1) or not (0) and other attributes that will support the classification results.

In this study, we propose a method which combines GBT and LR with Select by Weight in spambase dataset. The proposed model is expected to predict email including spam or not. Figure 1 shows the research process that will be carried out.
In accordance with the picture, we will do two experiments. The first experiment was to only use GBT while our second experiment implemented LR to handle noisy data before training in GBT to avoid overfitting.

The first experiment, the model is formed using only GBT. Before we start processing data using classification algorithms, the data was divided into 70% training data and 30% testing data. Then the model is formed with 10-fold validation using the GBT classification algorithm. The model has been formed will be tested with 30% testing data. 70% of training data will be selected to reduce noisy data before entered into GBT.

The second experiment, we divided the data into 70% training data and 30% testing data. We added preprocessing data before forming the model using LR with Select by Weight. The training data was divided into 10 parts using 10-fold validation where 9 parts of data will be used as training data and the other parts will be used as testing data until 10 iterations. Then, the model will be formed using the GBT algorithm. After the model is formed, 30% of testing data was tested on existing models.

The method of our study will be evaluated using classifier effectiveness. The experiment result will produce a confusion matrix such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

| Actual Class (Observation) | Y   | N   |
|----------------------------|-----|-----|
| Predicted Class Expectation|     |     |
| Y                          | TP  | TN  |
| N                          | FP  | FN  |
The confusion matrix will be measured for calculation. Based on measurement calculations, the calculation as follows:

(i) Accuracy: percentage results from the measurement of samples that have been correctly classified.

\[ \text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \]

(ii) Sensitivity: percentage results from measuring true values as positive or TP.

\[ \text{SN} = \frac{TP}{TP + FN} \]

(iii) Specificity: percentage results from measuring true values are negative or TN.

\[ \text{SP} = \frac{TN}{TN + FP} \]

4. Experimental Results

The research has been done using platform based on Intel(R) Celeron(R) CPU N2840 @ 2.16GHz 2.16 GHz, RAM 2GB, Windows 7 64-bit as Operating System and Rapidminer version 8.1.0.0 as a data analysis tool. Using rapidminer can be produce final calculations such as accuracy and confusion matrix.

First of all, we will show our experimental results using only GBT. The number of trees has been inputted in GBT is 55, maximal depth 5, min rows 10.0, min split improvement 0.0, number of bins 20, learning rate 0.1, and sample rate 1.0. The results show that GBT can obtain 95.03% accuracy with a confusion matrix like the following table.

| Pred. 1 | True 1 | True 0 | Class precision |
|---------|--------|--------|-----------------|
| Pred. 1 | 517    | 20     | 96.28%          |
| Pred. 0 | 27     | 402    | 93.71%          |
| Class recall | 95.04% | 95.26% |

Next, we will add preprocessing using LR with Select by Weight. We input the weight relation such as all but bottom k with a value of k is 5. The result has obtained from the second experiment. The results showed an increase in accuracy of 0.10% from the previous experiment.

| Pred. 1 | True 1 | True 0 | Class precision |
|---------|--------|--------|-----------------|
| Pred. 1 | 517    | 20     | 96.28%          |
| Pred. 0 | 27     | 402    | 93.71%          |
| Class recall | 95.04% | 95.26% |

From the two experiments, it showed that the addition of LR preprocessing with Select by Weight can improve the performance of GBT by 0.10%. This proves that increasing LR with Select by Weight to GBT gives better results than using only GBT. We conclude the comparison of calculated results from two models that have been formed.
Finally, the research we conducted by proposing LR with Select by Weight and GBT resulted in better accuracy than previous studies. The comparisons of accuracy obtained from previous studies and using the same dataset, can be seen in the following table.

| Year | Method | Result |
|------|--------|--------|
| 2014 | Cost-sensitive three-way decision | 89.88% |
| 2014 | NSA-DE | 83.06% |
| 2014 | NSA-PSO | 91.22% |
| 2015 | NSA-PSO | 83.20% |
| 2016 | LRFNT+DT | 91.67% |
| 2017 | Fuzzy Granular Classifier | 94.62% |
| 2018 | **Proposed Method** (LR with Select by Weight+GBT) | **95.13%** |

5. Conclusion and Future Work

The experiments have been carried out in this research have given a good result. The additions of pre-processing using LR with Select by Weight can improve the performance of GBT as big as 0.10%. In this study, we are expected to be able to produce spam filtering that can detect spam emails. The results of accuracy obtained can be increased again by adding another process or replacing the models has been used before.

In the future study, we expect the proposed method can be produce better accuracy so detecting spam emails will be more effective.

References

[1] Tuteja, S.K. and N. Bogiri. *Email Spam filtering using BPNN classification algorithm*. in Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. 2016. IEEE.

[2] Lee, S.M., et al. *Spam detection using feature selection and parameters optimization*. in Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on. 2010. IEEE.

[3] Ergin, S. and S. Isik. *The assessment of feature selection methods on agglutinative language for spam email detection: A special case for Turkish*. in 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings. 2014.

[4] Gomes, S.R., et al. *A comparative approach to email classification using Naïve Bayes classifier and hidden Markov model*. in Advances in Electrical Engineering (ICAEE), 2017 4th International Conference on. 2017. IEEE.
[5] Zhou, B., M. Lu, and Y. Wang. *Counting people using gradient boosted trees*. in *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*. 2016. IEEE.

[6] James, G., et al., *An introduction to statistical learning*. Vol. 112. 2013: Springer.

[7] Bhargava, K. and R. Katarya. *An improved lexicon using logistic regression for sentiment analysis*. in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*. 2017.

[8] Wei, D., T. Wang, and J. Wang, *A logistic regression model for Semantic Web service matchmaking*. Science China Information Sciences, 2012. 55(7): p. 1715-1720.

[9] Wijaya, A. and A. Bisri. *Hybrid decision tree and logistic regression classifier for email spam detection*. in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*. 2016.

[10] Hastuti, K., *Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif*. Semantik, 2012. 2(1).

[11] Zhou, B., Y. Yao, and J. Luo, *Cost-sensitive three-way email spam filtering*. Journal of Intelligent Information Systems, 2014. 42(1): p. 19-45.

[12] Idris, I., A. Selamat, and S. Omatu, *Hybrid email spam detection model with negative selection algorithm and differential evolution*. Engineering Applications of Artificial Intelligence, 2014. 28: p. 97-110.

[13] Idris, I. and A. Selamat, *Improved email spam detection model with negative selection algorithm and particle swarm optimization*. Applied Soft Computing, 2014. 22: p. 11-27.

[14] Idris, I., et al., *A combined negative selection algorithm–particle swarm optimization for an email spam detection system*. Engineering Applications of Artificial Intelligence, 2015. 39: p. 33-44.

[15] Salehi, S., et al., *Fuzzy granular classifier approach for spam detection*. Journal of Intelligent & Fuzzy Systems, 2017. 32(2): p. 1355-1363.

[16] Bache, K. and M. Lichman, *UCI Machine Learning Repository (http://archive. ics. uci. edu/ml), University of California, School of Information and Computer Science*. Irvine, CA, 2013.

[17] Cranor Lorrie F. and A. LaMacchia Brian, *Spam*. Communications of the ACM, 1998(41): p. 8.