Algorithm Comparison on Email Spam Filtering Task

Yixuan Li *

Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States

* Corresponding Author Email: yxli@seas.upenn.edu

Abstract. Email has long been a major form of communication among organizations and individual users. In recent years, with the rise of internet use, email spamming has become increasingly common. Spamming has raised security concerns as it causes potential loss to the users with fake advertisements, invalid information, undetected virus, and other harmful information. Various techniques have been developed to facilitate spam filtering, using classifying algorithms to characterize emails into different categories. This article presents an investigation of how machine-learning-based algorithms are used in email spam filtering by providing some previous researches that have shown to be successful. The algorithms are in the range from supervised learning, including the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) to unsupervised learning such as artificial neural networks (ANN) and partitional clustering. This article also presents an experiment that compares different implementations of such algorithms, including the SVM, NB, and K-Nearest Neighbor (KNN). The results reflect that NB gave the highest accuracy. A second attempt of the same experiment was conducted, with an improved data cleaning procedure and larger testing sets. The data collected from the second attempt again show that the NB implementation gave the highest accuracy in detecting spam emails.

Keywords: Spam filtering, machine learning, email spam classification, Support Vector Machine, Naïve Bayes, KNN.

1. Introduction

Machine learning is a tool that facilitates data analysis on complex data sets and enables accurate model building [1]. The goal of machine learning is to reduce human intervention in the computational process by making the machine learn from a training set and automatically output a prediction or a detected pattern.

As email spamming becomes growingly frequent, researchers have been developing spam filtering algorithms and exploring the utilization of machine learning in spam filtering. Successful experiments have been conducted to design email filters with relatively high accuracy rates. This section discusses the current status of research on email spam filtering and introduces several commonly used machine learning models.

This article introduces the background of the email spamming issue. It then gives an overview of previous work dedicated to designing spam filtering algorithms. The methodology is presented, which contains an introduction to the data set, the data preprocessing procedure, an introduction to the models implemented in the experiment, and the optimization procedure. The results are then presented, followed by a discussion on potential improvements in the future study.

2. Related Work

There are primarily two different techniques to machine learning: supervised machine learning and unsupervised machine learning. The primary distinction between the two is the kind of data they incorporate. Labeled data are the types of data that are used in the supervised machine learning models' training processes. The models then output a prediction, which can either be a classification or a regression [2]. On the other hand, unsupervised machine learning takes in unlabeled data and uncovers the patterns of the dataset through clustering or dimensionality reduction [3]. Figure 1

depicts the rough structure of different types of machine learning and the major algorithms under each category.

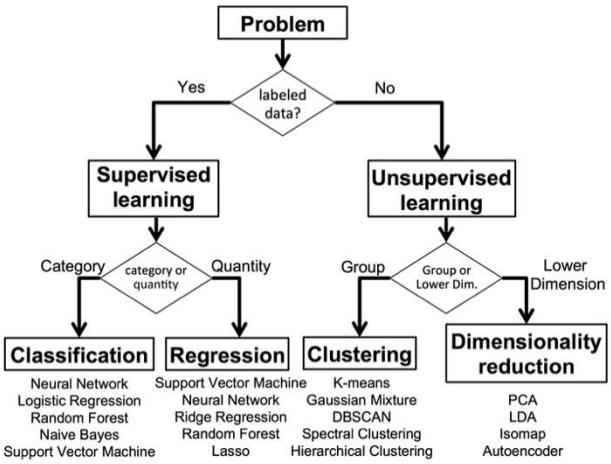


Figure 1. Types of Machine Learning [4]

Some of the major machine learning algorithms for spam filtering include the SVM, NB Classifier, decision tree, ANN, KNN, and unsupervised clustering methods such as partitional clustering [5]. SVM is a powerful tool that has high classification accuracy. It functions by outputting a hyperplane from the training data, achieving classification by separating data into different categories using the hyperplanes [6]. NB is a probabilistic classifier that assumes "absolute independence of features" [7]. In other words, it assumes that one specific feature of a class has no relation to other features of the class. Decision trees maintain a tree structure by recursively calling binary splitting according to a parameter. The nodes of the tree are the splitting points, and the leaves are the final outcomes [8]. One important type of DT method is the random forests, which produce a multitude of decision trees. ANN algorithms are tools that mimic how human brains process information and are based on the structure of a human neuron. It often consists of multiple layers to carry out the computational operations [9]. The KNN method assigns data points to their nearest neighbor by calculating the minimum distance between the data points. It has the advantage of a short training stage, with the sacrifice of a potentially costly testing stage [10]. Partitional clustering is an unsupervised approach. It separates the data set into partitions, where each partition represents a disjoint cluster [11]. Research had been initiated in the past to investigate the different spam filter implementations. Bandy and Jan had conducted a comprehensive review on statistical spam filter techniques, covering models such as NB, SVM, Term Frequency-Inverse Document Frequency (TF-IDF), Bayesian Additive Regression Trees (BART), and KNN [12]. Cormack explored and tested the efficiency of a variety of filtering methods developed up until 2008 [13].

3. Methodology

There are primarily two different techniques to machine learning: supervised machine learning and unsupervised machine learning. The primary distinction between the two is the kind of data they incorporate. Labeled data are the types of data that are used in the supervised machine learning models' training processes. The models then output a prediction, which can either be a classification or a regression [2]. On the other hand, unsupervised machine learning takes in unlabeled data and uncovers the patterns of the dataset through clustering or dimensionality reduction [3]. Figure 1 depicts the rough structure of different types of machine learning and the major algorithms under each category.

3.1. Data

The dataset used for this experiment is from Kaggle [14]. It is a CSV file containing 5728 rows and 2 columns, which are 'text' and 'spam.' The 'text' columns contain the email headers and

contents, and the 'spam' column uses 0 and 1 to indicate whether an email is a spam or ham (non-spam). Figure 2 gives a general visualization of the ratio of spams and hams in the data set.

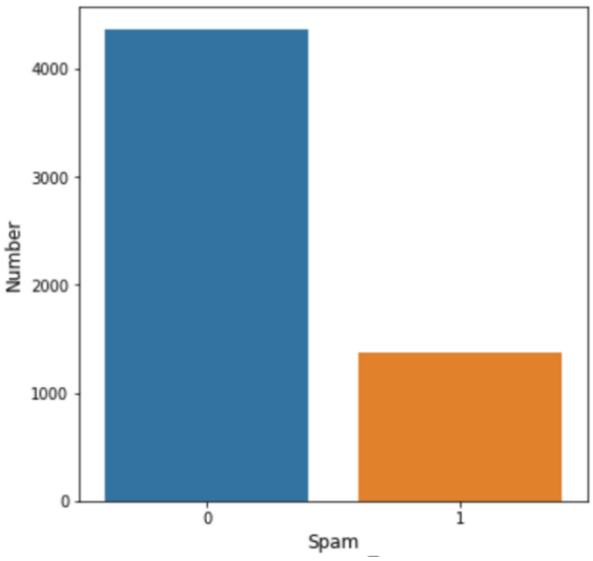


Figure 2. Number of Spam and Ham in the Data Set

3.2. Data preprocessing

The pre-processing of this dataset utilizes Natural Language Processing (NLP). It first removes all punctuations and English stop words, which are the most commonly used English words that do not necessarily have important meanings in most emails. Removing the stop words helps decrease the size of the dataset and prevent the intervention of the frequently appearing words that do not have strong indications of the actual content of an email. Then, stemming is applied to each word. Stemming is the procedure that reduces the inflectional form of a word to its root or stem [15]. This is an efficient and useful way of normalizing the text.

Then, the frequency matrix is generated from the bag of words and this vector can now be employed to fit the classifiers. The dataset is split into two parts, one of which is the training set and the other the testing set. The former is utilized to implement classification models, and the latter is used to calculate the accuracy rates of different implementations.

3.3. Model

The experiment gives a general overview of the raw data, pre-processes the dataset by cleaning the data using natural language processing (NLP), and applies the SVM, Naïve Bayes, and Random Forest models individually from the Python scikit-learn (sklearn) library to produce a classification output.

A general exploratory of the raw data shows that the dataset contains 5728 entries, all of which are non-null. Furthermore, there are 4360 non-spam emails and 1368 spam emails.

3.3.1 Support vector machine

The SVM is often employed in two-group classification problems. It falls under the category of supervised machine learning, meaning that it trains using labeled data. SVM achieves classification by generating a hyperplane called the decision boundary, which separates the two groups to the two sides of the hyperplane. This model is well suited for text classification problems.

3.3.2 Naïve Bayes

The NB model is developed based on the Bayes' Theorem, which is helpful for determining the probability of one event happening given that another event has happened. Bayes' Theorem assumes that all predictors are independent. In other words. The predictor has no impact on the probability of

the existence of other predictors. It also assumes that all the predictors are equally influential to the result. The NB model has shown good performances in NLP problems.

3.3.3 Random Forest

The random forest is employed in classification and regression algorithms, and it belongs to the category of supervised learning machine. A forest is composed of decision trees. Each tree gives a classification and contributes to making the final classification by means of voting. Random forest algorithms are efficient in processing large data sets and can give indications of the importance of the different features.

3.3.4 K-Nearest Neighbors

The KNN method is another supervised learning algorithm that is useful for building classification and regression models. KNN model assumes that the proximity in distance is an indication of similarity among objects. It computes the distances between data points and chooses the k nearest points. For classification problems, it generates the majority voting of the K labels.

3.3.5 Optimization

In order to optimize the results, another experiment is performed in the further step. The dataset can be better cleaned by dropping more unnecessary elements, which can also help shorten the running time. There should also be accuracy tests based on larger testing sets. Furthermore, clustering approaches can be incorporated as another alternative.

A second experiment is conducted following the first experiment in an attempt to improve the performance, addressing the insufficiencies of the first experiment. This experiment follows the same steps that the first experiment adopted but made slight changes for each step. For data pre-processing, additional steps to drop duplicated values and to remove the common starting phrase “Subject” were employed, which can reduce the running time by eliminating the interferences from repeated meaningless entries. This attempt also modifies the testing set ratio from 20% to 40% to obtain larger testing sets. Lastly, the KNN model is added to the implementation for a more comprehensive comparison.

4. Results

The results obtained from the first experiment using implementations of SVM, random forest, and NB are listed in Table 1. The first implementation uses the SVM model, which gives an accuracy score of about 98.2%. The second implementation uses the random forest model, which gives an accuracy score of about 98.3%. The third implementation uses the Naïve Bayes model, which gives an accuracy score of about 99.2%. According to the accuracy scores obtained from the above implementations, the NB model achieves the highest accuracy. This result indicates that the NB model is potentially one of the most suitable models for spam filtering, using the current data pre-processing procedure and the current programming structure.

Table 1. Results from the First Experiment

Model	Test Size	Accuracy
SVM	0.2	98.1675392670157%
Random Forest	0.2	98.34205933682374%
NB	0.2	99.21465968586387%

Table 2 illustrates the data collected from the second experiment. In this attempt, SVM, random forest, and NB implementations are tested again, and the implementation of the KNN model is added for comparison. The NB model continues to maintain the highest accuracy score, which is about 99.3%. Random forests and SVM had similar performances, both giving an accuracy of about 97%. The KNN model shows the worst accuracy score of only 87.8%. The Naïve Bayes indeed seems to

have the most efficient and accurate performance. Out of all the implementations used by this experiment, it consistently outputs highly accurate results.

Table 2. Results from the Second Experiment

Model	Test Size	Accuracy
SVM	0.4	97.54170324846356%
Random Forest	0.4	97.41000877963125%
NB	0.4	99.29762949956101%
KNN	0.4	87.84021071115014%

5. Discussion

There are still potential improvements that can be made from the second improvement. Further research needs to be done on tokenization and lemmatization. Lemmatization might be a good replacement for stemming, which, although powerful, could result in over-stemming and thus the elimination of meaningful words. Furthermore, more testing trials with varied training-test splits can be added to reflect the consistency of the results. There is also room for comparing unsupervised learning models, such as the partitional clustering approach, which might provide a new perspective.

Some questions can also be raised for the future study. It is worth considering if it would be possible to further divide the emails into subcategories. For instance, spam emails can be classified as advertisements, promotions, invalid information, harmful links, and etc. Another potential advancement is to accomplish email filtering by only taking in the headers of emails, which will save a large amount of time and prevent information from leaking.

6. Conclusions

As email spamming continues to exist as a risk-posing factor to information security for organizations and individual users, machine learning presents a promising solution to accurate and efficient detection of spam emails. This article introduces how the supervised and unsupervised learning have been used in spam filtering by delivering an overview of some important past research on the utilization of machine-learning algorithms, including the SVM, NB, ANN, and partitional clustering. To test which implementation gives the most accurate result, an experiment was conducted to compare the SVM, NB, and KNN methods. The pre-processing of data makes use of NLP, including punctuation removal, stop words removal, and stemming. Then, using the sklearn library, the algorithms mentioned above were tested, and the implementation using Naïve Bayes achieved the highest accuracy. To improve the first experiment, a second experiment was conducted to incorporate a better data cleaning procedure and utilized larger testing sets. According to the results obtained from the second experiment, the Naïve Bayes algorithm can achieve the highest accuracy.

References

[1] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.

[2] Jiang, T., Gradus, J. L., &Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. Behavior Therapy, 51(5), 675–687.

[3] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., &Aljaaf, A. J. (2019). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Unsupervised and Semi-Supervised Learning, 3–21.

[4] Bayen, T. (2020). Python Programming and Numerical Methods: A Guide for Engineers and Scientists. S.L.: Elsevier Academic Press.

[5] Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges. Security and Communication Networks, 2022, 1–19.

[6] Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567.

[7] Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Technical Note: Naive Bayes for Regression. Machine Learning, 41(1), 5–25.

[8] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics, 18(6), 275–285.

[9] Shi, T., & Horvath, S. (2006). Unsupervised Learning With Random Forest Predictors. Journal of Computational and Graphical Statistics, 15(1), 118–138.

[10] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, 986–996.

[11] Celebi, M. E. (Ed.). (2015). Partitional Clustering Algorithms. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-09259-1

[12] Banday, M. T., & Jan, T. R. (2009). Effectiveness and limitations of statistical spam filters. https://arxiv.org/abs/0910.2540

[13] Cormack, G. V. (2008). Email Spam Filtering: A Systematic Review. Foundations and Trends® in Information Retrieval, 1(4), 335–455.

[14] Kaggle. 2017. Spam filter. https://www.kaggle.com/datasets/karthickveerakumar/spam-filter?resource=download

[15] Srivastava, T. 2019. NLP: A quick guide to Stemming. https://medium.com/@tusharsri/nlp-a-quick-guide-to-stemming-60f1ca5db49e