CS5262 Final Report

Yizhou Guo

1 Introduction

Email spam is a well-researched topic in machine learning. Researchers have applied multiple learning methods, unsupervised, semi-supervised and supervised, to various dataset including the dataset my am using, Enron-Spam. Among the methods used, some are more popular than the others, including SVM, k-NN, Genetic Algorithm(GA), ANN, Naïve Bayes, MLP, and Random Forest.[1,2]

Before going deeper into the topic, it is necessary to know the basics of email spam. Usually, email spams are to advertise (legit or illegal), bait for a fraud scheme (phishing), promote a cause, or spread computer malware.[3] Modern email service providers like Google(Gmail), Microsoft(Outlook), and Apple(iCloud) further categorize spam into advertisement, social, and junk mails. This requires a lot efforts on whitelisting and blacklisting and is not include in my project.

Among the datasets researched, there are mainly three types: raw email, email-body only, and pre-processed then tokenized. While tokenized data are easier to process and to learn, it contains less information, including some specification we want to know. Raw email contains information such as sender address, title, and time sent. I am not building a comprehensive email filter, so the extra information is not to my interest. For the reasons above, my project focus on email-body only type of dataset.

As a result of the difference in dataset, the result vastly differs, especially when using tokenized dataset. In previous research, highest accuracy reached

99.87% using ANN[4], and highest accuracy of raw email or email-body only dataset reached 99% using MLP[5]. Other research have accuracy range in 76% to 99%.[1] Of the researchers using the same dataset, Enron-Spam, the accuracy range in 85%(using random forest) to 98.76%(using Deep Neural Networks).

2 Dataset & Methods

The dataset I settled on is Enron-Spam 1, and I am applying Multiple Linear Regression, SVM, ANN, and CNN to train the data.

3 Feature Selection

Feature selection is a crucial step in applying machine learning methods to email-body only dataset. To process the data, I used 5000 most commonly used English words and count the number of occurrences of each word, to tokenize the data. Beyond that, special character count, unrecognized words count, as well as re-recognized words (after removing non a-z characters) are also included in the features. These features are included because there are plenty of character look-alikes in Unicode, and spam email senders tend to use these characters to fake the email detection system. Also, some spam email senders like to use special characters like $$$ to represent some key words, in this case, money.

After the data is tokenized, feature must be further filtered due to computational power restrictions. To select the features, I used most significant occurrence rate difference and most significant occurrence rate ratio for words appeared more than 5 times. Surprisingly, the occurrence rate ratio outperforms difference, especially when using small number of features.

4 Results

Chart 1: Different features in MLR (accuracy):

10 difference	10 ratio	10 ratio + 3 engineered	10 difference + 10 ratio + 3 engineered
71.9%	72.5%	75.6%	75.8%

Chart 2: Different threshold in MLR (20+3 features):

Threshold	Accuracy	False Positive Rate
0.5	77.9%	20.5%
0.6	75.9%	12.5%

Chart 3: Different kernal in SVM (20+3 features):

Kernal	Accuracy	False Positive Rate
Linear	78.7%	23.8%
Poly	72.6%	22.2%
RBF	79.5%	18.6%

Chart 4: Different layers and cells in ANN (threshold=0.5):

(Layer, Cells))	Accuracy	False Positive Rate
(2, [512,256])	85.7%	20.7%
(2, [128,64])	85.2%	21.3%
(1, [512])	84.5%	18.3%

Chart 5: Different layers and cells in ANN (threshold=0.6):

(Layer, Cells))	Accuracy	False Positive Rate
(2, [512,256])	86.8%	17.2%
(2, [128,64])	85.6%	17.5%
(1, [512])	84.9%	20.1%

Chart 6: Different learning methods using 20+3 features:

Method	Accuracy	False Positive Rate
MLR	75.8%	12.5%
SVM(RBF Kernal)	79.6%	23.8%
ANN	84.6%	25.5%
CNN	74.9%	25.0%

Chart 7: Different learning methods using 800+3 features:

Method	Accuracy	False Positive Rate
MLR	86.7%	7.9%
SVM(RBF Kernal)	93.2%	8.4%
ANN	95.6%	7.7%
CNN	70.9%	50.0%

5 Discussion and Conclusions

From chart 1, it is clear that the three engineered features plays an important role in elevating prediction accuracy. From chart 2, we know false positive rate may significantly decrease when using higher threshold and lower accuracy as trade-off. From Chart 3, we know selection of kernals largely affects the model performance, and RBF kernal performs the best to the data. From Chart 4 and 5, we found two layers with [512, 256] as number of cells and 0.6 as threshold perform the best in ANN.

Chart 6 and 7 are complete results of different methods. Number of features largely affects model performance. However, it does require a lot more time to train a model with 800+3 features. ANN appears to be the best model among the 4 compared, both in accuracy and false positive rate (when using 800+3 features). The results, especially ANN model, are effective, in comparison to the research of same dataset.

6 Future possibilities

NLP and LLM models can be applied to the dataset, given the nature that the dataset is in text form. Also, OCR technology may also be applied to email spam detection, given spam email senders are adapting to the current email filtering technologies.

7 GitHub Repository

https://github.com/guoyizhou01/EmailSpamDetection

8 Reference

[1] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019

[2] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, Machine learning for email spam filtering: review, approaches and open research problems, Heliyon, Volume 5, Issue 6, 2019

[3] Gordon V. Cormack (2008), "Email Spam Filtering: A Systematic Review", Foundations and Trends® in Information Retrieval: Vol. 1: No. 4, pp 335-455.

[4] Nossier, Ann & Nagaty, Khaled & Taj-Eddin, Islam. (2013). Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks. International Journal of Computer Science Issues. 10. 17-21.

[5] Ali Shafigh Aski, Navid Khalilzadeh Sourati, Proposed efficient algorithm to filter spam using machine learning techniques, Pacific Science Review A: Natural Science and Engineering, Volume 18, Issue 2, 2016, Pages 145-149,