Fedir Prokhnytskyi, Oleksandr Rokovyi.

MAIL MESSAGE FILTERING BASED ON ARTIFICIAL INTELLIGENCE

Abstract. The purpose of the work is to analyze the effectiveness of the email filtering module. The research uses a dataset from the Kaggle platform that has been processed and supplemented with additional messages, as well as classifiers based on two different models: a naive Bayesian classifier; support vector machines method. The effectiveness of each of the approaches based on the same sample was analyzed using model metrics: precision and recall. Each model was further tested on the test data set.

Keywords: artificial intelligence, machine learning, classifier, neural network.

Introduction

The increase in the number of unsolicited emails, called spam, has created a need for spam filters to reduce the time and effort involved in managing inboxes as well as managing storage. Spam does not allow the user to fully and effectively use time, memory and network bandwidth. The sheer volume of spam flowing through computer networks wreaks havoc on mail servers' memory space, communication bandwidth, processor power, and user time. Effective spam filters can prevent cyber fraud to users and data can also be protected from spammers. Recently, machine learning techniques have been extremely successful in detecting and filtering spam. These models mainly "learn" on data sets that are previously formed and are able to find "commonality" in new data that comes from the outside.

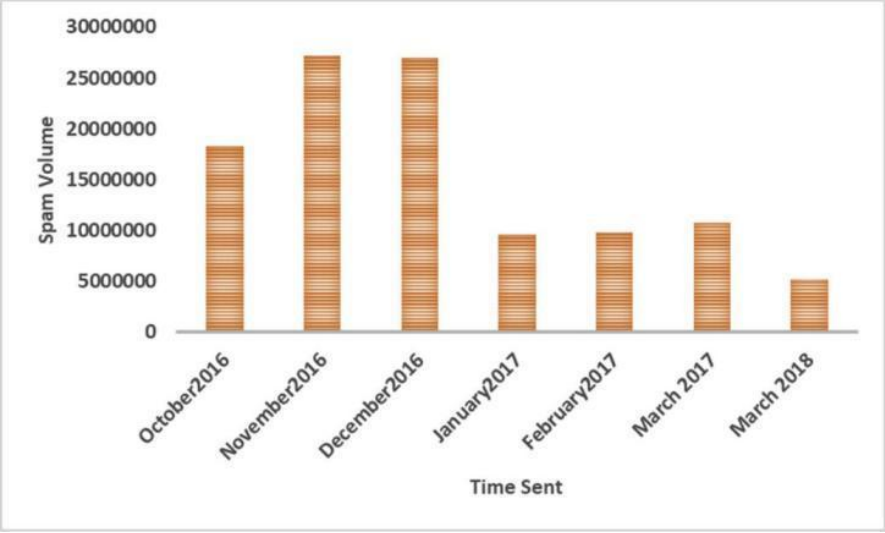

Fig. 1. Volume of spam from 2016 to 2018

According to Kaspersky Lab's report, in 2015 the volume of spam was down to a 12-year low. Spam volume fell below 50% for the first time since 2003. According to anti-virus software developer Symantec, spam fell to 49.7% in June 2015 and 46.4% in July 2015. The latest statistics from 2016 show that spam accounted for 56.87% of email traffic worldwide, with the most prominent types of spam being medical and dating spam.

Analysis of existing solutions

The Naive Bayes classifier is a supervised learning method based on probability and statics. This method of filtering letters uses an adaptive set of rules, and the corresponding set of probabilities is set according to the classification decisions and received letters. Each mail is described by a set of attributes, and each attribute is assigned a probability according to the number of times it occurred in the training set. A naive Bayes classifier for spam filtering uses a simple probability formula that can be interpreted as (where c=spam): "The probability that a message will be spam, given its characteristics, is equal to the probability that any message will be detected as spam, multiplied by the probability of features occurring in spam, divided by the probability of detecting those features in any message."[1]

The advantage of this approach is that the sample size requirements are reduced from exponential to linear. The disadvantage is that the model is accurate only if the independence assumption holds.

The support vector machines method is a supervised learning algorithm that has shown much better performance than other classifiers due to its multidimensional bounds and simplicity. It maximizes the distance to the nearest reference point, and points equidistant from a given reference point are called support vectors. A linear combination of these support vectors forms a classifier or partition hyperplane.

Algorithm: the training set S is introduced, and the kernel function is determined in the form {c1, c2,…..,cn} and {d1, d2,…..,dn}. The number of nearest neighbors, say k, is assigned. Then a two-stage for loop is designed, c=ci from 1 to n is set for the outer loop. The inner loop is performed for j from 1 to q, where the SVM classifier function f(x) is designed with the fusion parameters (c, d). Using the if-else condition, the classifier function f(x) is compared with the best classifier given by the k-fold cross-validator. Therefore, a return command is given to classify the message as spam or non-spam, shown in Figure 2.

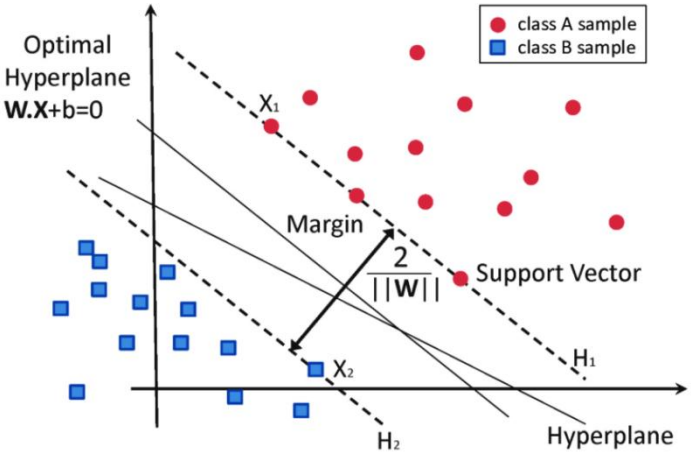

Fig. 2. Support vectors separate elements from different classes (spam/non-spam)

Description of the features of the work results

The assessment of the relevance of the models is based on the comparison of the metrics of these models. Two metrics will be used in this study: precision and recall. Precision (also known as positive predictive value) is the proportion of relevant instances among retrieved instances, while recall (also known as sensitivity) is the proportion of relevant instances that were retrieved. Therefore, both precision and recall are based on relevance.

Consider a computer program for recognizing dogs (the corresponding element) in a digital photograph. After processing an image containing ten cats and twelve dogs, the program identifies eight dogs. Of the eight items identified as dogs, only five are actually dogs (true positives), while the other three are cats (false positives). Seven dogs were missed (false negatives) and seven cats were correctly excluded (true negatives). The program's precision is then 5/8 (true positives / selected items) and its recall is 5/12 (true positives / matched items). When a search engine returns 30 pages, only 20 of which are relevant, instead of returning 40 additional relevant pages, its precision is 20/30 = 2/3, which tells us how valid the results are, while its recall is 20/ 60 = 1/3, which tells us how complete the results are. Adopting a statistical hypothesis testing approach in which the null hypothesis in this case is that the subject is irrelevant, i.e. not a dog, no Type I and Type II errors (i.e., perfect specificity and sensitivity of 100% each) corresponds to perfect accuracy, respectively (no false positives) and perfect recall (no false negatives). In general, recall is simply the addition of the Type II error rate, that is, one minus the Type II error rate. Accuracy is

related to the Type I error rate, but in a slightly more complicated way because it also depends on the prior distribution of viewing the relevant and irrelevant item.

The cat and dog example above contained 8 − 5 = 3 Type I errors, for a Type I error rate of 3/8, and 12 − 5 = 7 Type II errors, for a Type II error rate of 7/12. Precision can be thought of as a measure of quality and recall as a measure of quantity. Higher precision means that the algorithm returns more relevant results than irrelevant ones, and high recall means that the algorithm returns most of the relevant results (whether or not irrelevant results are also returned). In fig. 3.1 shows the distribution of false positive/negative values and formulas for calculating accuracy and recall.[2]

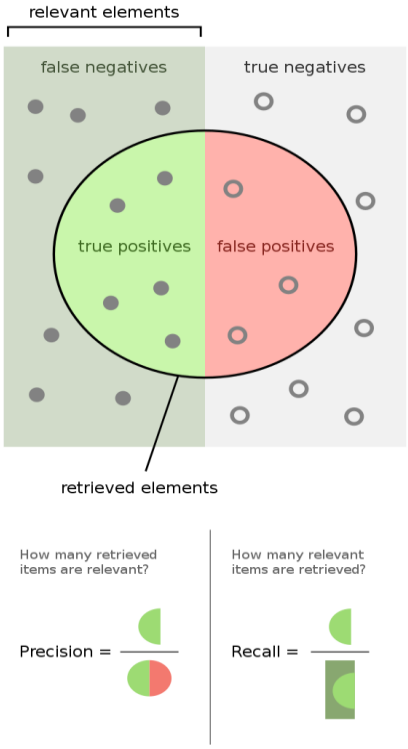

Fig. 3. The set of results of machine learning models

Also, a data set is specially selected for model training. It is not necessary that this set should contain the same amount of data of both classes (spam, useful mail). In this study, model training takes place taking into account that the model should not contain false positive values, accordingly, it is much worse to misclassify useful mail: add it to the spam box, than to misclassify spam - add it to the inbox. This study used a data set with the following ratio of spam to useful mail.

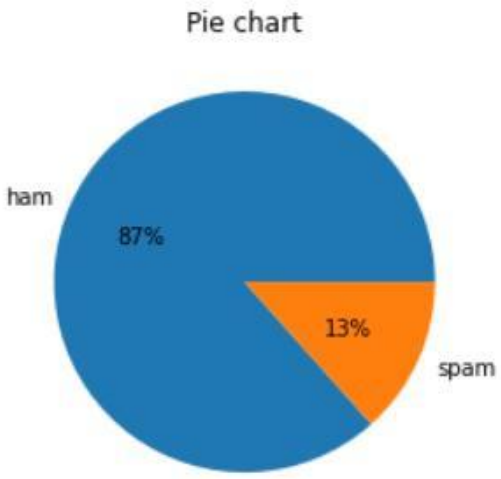

Fig. 4. The ratio of spam to useful mail in the data set

Comparison results of trained and tested models

The result of the study of models based on the Bayes classifier and based on the support vector machines method. Training and testing will be carried out, as well as metrics will be taken: accuracy, recall level, which will help to choose the best model. The best model will be the one whose metrics will be the highest, and also according to the results of testing, which will have the least false positive results, that is, such results when useful mail was determined as spam, while spam mail determined as useful is allowed, but nevertheless the model that will have fewer such results will still be better, that is, the goal of the study will be a certain compromise between the models with the least number of false positive results and the largest number of true negative results.

A model based on a naive Bayesian classifier

Many experiments with different hyperparameter alpha will be performed for this model. In machine learning, hyperparameter optimization or tuning is the problem of choosing the optimal set of hyperparameters for a training algorithm. A hyperparameter is a parameter whose value is used to control the learning process. In contrast, the values of other parameters (as a rule, node weights) are studied.[3]

In this study, the alpha parameter will be randomly selected from 0.00001 to 20 in steps of 0.11. Also, in the course of training, lists with precision and recall will be formed for each iteration, which will allow choosing the best iteration.

```
[12]: alpha              12.430010
      Train Accuracy      0.973205
      Test Accuracy       0.970636
      Test Recall         0.785124
      Test Precision      0.989583
      Name: 113, dtype: float64
```

Fig.5. The best model selected on the basis of metrics is built on the basis of the Bayesian classifier

The selected model has an accuracy index of 0.989 and a recall value of 0.785. Based on this model, you can build a table of spam/non-spam detection results.

	True result	False result
Ham	1592	5
Spam	43	199

Fig. 6. The result of model testing

As a result of testing, you can see that the model has five false-positive results and 43 true-negative results (spam emails that got to the spam box)

A model based on support vector machines methods

For this model, many experiments will be conducted with different hyperparameter C. In this study, the parameter C will be chosen randomly from 500 to 1000 with a step of 100. Also, during training, precision and recall lists will be formed for each iteration, which will allow to choose the best iteration. The selected model has an accuracy index of 0.994 and a recall value of 0.8099. Based on this model, you can build a table of spam/non-spam detection results.

```
C                  500.000000
Train Accuracy       1.000000
Test Accuracy        0.974443
Test Recall          0.809917
Test Precision       0.994924
Name: 0, dtype: float64
```

	True result	False result
Ham	1596	1
Spam	46	196

Fig. 7. The result of training and testing the best model, which was selected on the basis of metrics and built by the SVM method

As a result of testing, you can see that the model has one false-positive result and 46 true-negatives (spam emails that got to the spam box).

Conclusion

As a result of studying and testing two models based on different types of machine learning algorithms and based on the selected dataset, it was determined that the most usable model is the model based on the support vector machines method, because it has the highest accuracy, the highest recall and high speed, which makes it suitable for use on large data sets. This model is used as the core of the mail filtering system.

REFERENCES

1. Bayesian Optimization in a Billion Dimensions via Random Embeddings https://jair.org/index.php/jair/article/view/10983

2. Fundamentals of Deep Learning http://perso.ens- lyon.fr/jacques.jayez/ Cours/Implicite/Fundamentals_of_Deep_L earning.pdf

3. Bergstra, James; Bengio, Yoshua (2012)."Random Search for Hyper-Parameter Optimization". Journal of Machine Learning Research. 13: 281–305. - https://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra1 2a.pdf

AUTHORS

Fedir Prokhnytskyi – PhD student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

Rokovyi Oleksandr – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.