UDC 004.772.:004.491 DOI: 10.30748/soi.2021.164.12

A. Shanin

Kharkiv National University of Radio Electronics, Kharkiv

ROTATION FOREST MODEL MODIFICATION WITHIN THE EMAIL SPAM CLASSIFICATION

Increased use of email in daily transactions for many businesses or general communication due to its cost-effectiveness has made emails vulnerable to attacks, including spam. Spam emails are unsolicited messages that are very similar to each other and sent to multiple recipients randomly. This study analyzes the Rotation Forest model and modifies it for spam classification problem. Also, the aim of this study is to create a better classifier. To improve classifier stability, the experiments were carried out on Enron spam, Ling spam, and SpamAssasin datasets and evaluated for accuracy, f-measure, precision, and recall.

Keywords: *Email spam classification, Rotation Forest, Naïve Bayes, Particle Swarm Optimisation, PV-DM, TF-IDF.*

Introduction

General Problem Statement. Email is a extremely fast and cost-effective mean of transmitting information from anywhere in the world, which can be used from personal computers, smartphones and other electronic gadgets of the latest generation [1].

Despite the increasing use of other forms of Internet communication, such as messaging and social networks, emails continue to occupy a leading position in business communications and is still necessary for other forms of communication and transactions.

Almost all people use email. In 2021, the number of email accounts worldwide was estimated at 4 billion, which is more than half of the world's population [2].

The growing popularity and use of emails for transactions has led to an increase in spam worldwide. Spam emails are unsolicited messages sent by email to several recipients who did not wish to receive these messages. The spammer has no previous relationship with the recipients, but collects their addresses from various sources, such as phone books and completed forms. Since email messages are the main means of sending harmful information, including viruses and phishing attacks, the number of spam messages is growing rapidly and is one of the most serious threats to email users [7–8].

Many researchers are already working on spam filtering techniques, but accurate spam detection is considered a challenge for many reasons, including the subjective nature of spam, the overhead of processing and delaying messages, the type of language used, and dinamical cost of filtering errors. To classify mail into regular messages and spam text classification approach is used [3].

Classification or prediction tasks, which are solved by supervised learning, seek to reveal hidden associations between the target class and independent variables, are widely used in data extraction. For supervised learning, classifiers allow you to assign labels to observations so that unobserved data can be classified based on learning data. Spam detection systems are built using classification algorithms and group emails as spam or regular messages [4]. Therefore, for effective filtering we need to create a system that will use the most effective methods of spam classification.

Analysis of the Recent Research and Publications. In the last years the growing number of email users as well as growing number of email with spam have made the task of processing large volumes of e-mails a difficult task for data mining and machine learning. This has led a number of researchers to conduct comparative studies on the effectiveness of classification algorithms using a different performance metrics for solving spam classification task.

A number of recent publications [4–11] were considered, where compared the efficiency of some algorithms of email classification and their modifications. Different algorithms were developed and compared by different authors and on different data sets. In [4], the most extensive previous studies of other authors, the classification algorithms they used for comparison or modification, and some new models were analyzed. The following standard algorithms were considered: Naive Bayes, Logistic Regression, K-neighbors [5], ANN [5], SVC, Random Forest, Random Tree, J48, multilayer perceptron, SVM [9] and other less known or their modifications: C-PLS, C-RT, CS-CRT, CS-MC4, CS-SVC, SCS-SCM [6], Continuous PLS-DA, PLS-LDA, LDA [1], Bayesnet, Rotation Forest [4] , Bayesian Logistic Regression, Hidden Naïve Bayes, Voted Perceptron, REP Tree [4; 10–11].

Artificial neural network and particle swarm optimization (PSO) was combined with a support vector machine for spam classification and separation problem by [9]. Their algorithm was compared with Self Organizing Map and K-means which uses method of estimating the area under the curve (AUC). The results showed that this method is better than others.

In [10], conducted an experiment with many methods of spam classification, trying to find the most suitable classifier for separating email as spam and non-spam. Authors tested the effectiveness of many classifiers and found that in the analysis of the results the Naive Bayesian classifier (NB) provides an accuracy of 76%, which shows a result that is better than the other two classifiers, such as SVM and J48. Also classification and training time for the NB classifier is less than for other two ones, which means that the NB classifier is the best classifier among the other two for spam classification.

A study was conducted [15] where particle swarm optimization (PSO) was used along with a naive bayes algorithm. The study showed that PSO gives for Naive Bayes a significant increase in classification accuracy within the framework of this task.

In the work that showed the most capacious research [4], were compared more promising algorithms that were discovered earlier, such as Bayessian Logistic Regression, Hidden Naive Bayes, RBF Network, Voted Perceptron, Lazy Bayessian Rules, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Multilayes perceptron, Naïve Bayes, J48, Random Tree. A comparison of these algorithms showed that the Rotation Forest spam algorithm most effectively recognizes spam.

The Aim of the Research is the creation of the most effective spam classifier. Since Rotation Forest showed the most effective result in this problem, the purpose of this study was also determined to improve this algorithm.

Statement of basic materials

Algorithm concept

In this section the basic concepts of the classification algorithm created on the basis of Rotation Forest.

To improve it were consider the work with the algorithm Rotation Forest concept description [12]. In this paper options for improving the algorithm including the concept of using Rotation Forest as a basis for other ensemble models and the concept of using another algorithm as a basis for Rotation Forest such as the Naïve Bayes algorithm is discussed.

Ensemble methods usually use weak learning models as basic algorithms, i.e. those that do not differ much from conventional guessing. Rotation Forest is an ensemble model using bagging as the ensemble method and decision tree as the individual model. So Rotation Forest is a very strong model. That is most likely, the use of this algorithm as a basis for another ensemble algorithm will give a very slight increase as an algorithm in the classification of spam.

Considering the concept of using another algorithm as a basis for Rotation Forest, you can see that this may be a good idea. For example, many studies [3–4; 17; 20] have shown that the Naïve Bayes algorithm is the most efficient simple classifier for use in the spam classification problem, much more efficient than Decision Tree.

Considering various works using the naive Bayesian classifier as one of the basic algorithms for Rotation Forest [13–14], we can conclude that this modification is likely to be very effective.

In [15], we can see that the particle swarm optimization (PSO) [16] greatly increases the efficiency of finding the local minimum by the Bayesian classifier in spam recognition.

Based on these findings, Rotation Forest with Multinomial Naive Bayes as basic algorithm which will be trained using the PSO optimization is used in this study as the main classification algorithm. To define which modification affects better on the basic algorithm, we will compare different combinations of these modifications.

Classifier creation

The paper [17] describes the general process of creating a text-based spam classifier (Fig. 1).

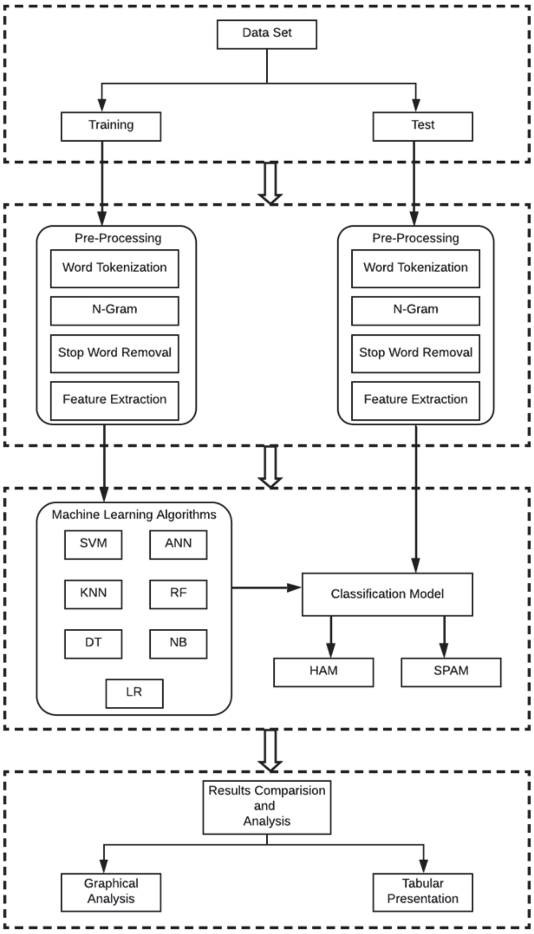


Fig. 1. The process of creating a text classifier
Source: described in the [17] paper.

Since we want to develop the most effective classifier, each part of the classifier must be chosen as best as possible. So we consider all parts of the classifier with more details:

1. Collection of datasets.

Based on the papers [17; 19], it was decided to take a set of datasets, which will be assembled into one, so that the classifier works more stably on unknown data. The following datasets were combined for the study:

– Enron spam (16545 spam and 17171 ham);

– SpamAssasin (1897 spam and 4150 ham);

– Lingspam (481 spam and 4212 ham).

The composite dataset of 18923 spam messages and 25533 regular messages was divided into 35565 messages for training and 8891 messages for testing algorithms.

2. Pre-processing of the dataset.

In order to work with text, the first we need to process it so that it is easy to vectorize. We will use the following methods of primary text processing, described in [18]:

– deleting links and digits or replacing them with a word;

– removal of punctuation marks, spaces;

– deletion of stop words, such as articles, prepositions, exclamations, etc.;

– stemming and lemmatization - transformation of a word into an initial form and infinitive;

– division of text into a set of tokens.

Following the recommendations [18] during preprocessing, we received sets of 1-gram and 2-gram, with which we will continue to work.

For further work we need to vectorize the text. There are many methods of vectorization of the text, but in [19] study was conducted and it shown that it is best to use a combination of two methods of vectorization in spam classification task: PV-DM and TF-IDF.

These methods complement each other – PV-DM has been trained to generate a vector for each word and each email and shows only the semantic meaning of words, while the TF-IDF method captures features that show high importance [20–21]. That is, the combination of these methods will analyze the semantics of words and their importance in the text.

3. Selection of algorithms for comparison.

The main goal is to compare the model Rotation Forest + Naive Bayes + PSO with other modifications. That is, it is Rotation Forest with Bayesian and Decision Tree (standard) with or without particle swarm optimization. The Random Forest model is also used for comparison, which is also based on the ensemble bagging method. Also we need to compare basic algorithms: Naive Bayes and Decision Tree.

Experimental results

The efficiency of the proposed algorithm is evaluated in terms of accuracy, f-measure, precision and recall. These parameters are calculated using next measures defined below [17]:

– True Positive (TP) – the number of emails with spam is correctly defined as spam;

– False Positive (FP) – the number of ham messages is incorrectly defined as spam;

– True Negative (TN) – the number of ham emails is correctly defined as ham;

– False Negative (FN) – the number of spam emails is incorrectly defined as ham messages.

Sets of experiments were performed, the results of which are displayed by the following metrics:

– Recall can be defined as the probability of correctly classifying spam. A high recall means that the filter tends to find as much spam as possible, no matter how many ham messages it identifies as spam. The formula is defined as follows:

$Recall = TP / (TP + FN)$.

– Precision defined as a fraction of correctly detected spam messages relative to all messages which detected as spam.

$Precision = TP / (TP + FP)$.

– Accuracy is the common ability of the filtering method to correctly classify spam and ham emails.

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$.

– F-measure: a popular indicator that combines precision and recall, calculating their mean harmonic value. This metric reflects the importance of spam classification only when the message is actually spam, rather than filtering all messages as spam. Defined as follows:

$F\text{-}measure = 2 * (Precision * Recall) / (Precision + Recall)$.

Fig. 2 shows the comparative results of algorithms that worked on data vectorized using the principles of PV-DM and TF-IDF [19] using the addition of 2-grams [18].

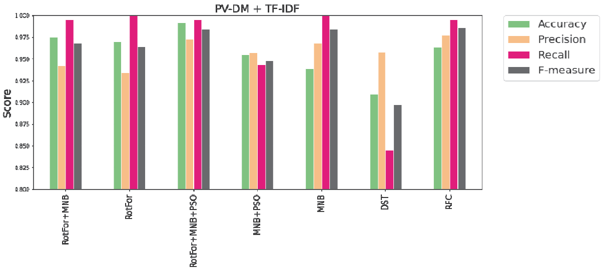


Fig. 2. The results of the algorithms according to the given metrics using PV-DM and TF-IDF

To be able to compare algorithms in standard conditions, experiments were performed without the use of 2-grams and using the following standard vectorization methods: TF-IDF (Fig. 3) and Bag of Words (BOW) (Fig. 4).

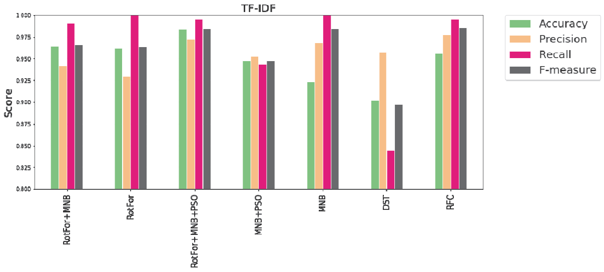


Fig. 3. The results of the algorithms according to the given metrics using TF-IDF

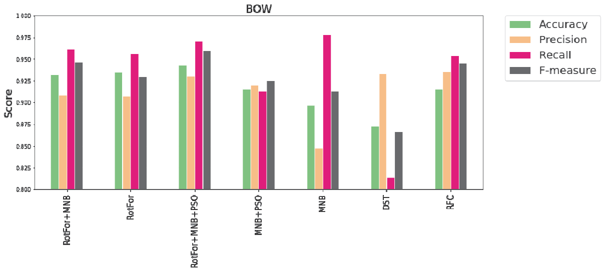


Fig. 4. The results of the algorithms according to the given metrics using BOW

ROC-curve is a graph that helps to assess the quality of a binary classification, as the classification result usually reflects the probability and the threshold of the classification result may change. Analysis of this curve provides an opportunity to rank models regardless of the costs context or class allocation. Fig. 5 shows the ROC curves of the classifiers used in this study.

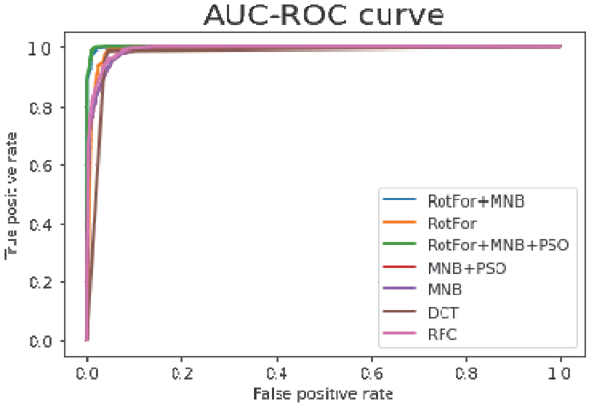


Fig. 5. Graph ROC (AUC) curve (PV-DM and TF-IDF)

Based on previous calculations, we can conclude that the proposed algorithm works much better than other algorithms selected for comparison, such as the standard Rotation Forest algorithm and variations of its modifications (Fig.5). The accuracy of classification of the proposed algorithm increases all other algorithms by at least 1.7%. The main advantage of this algorithm is the use of PSO optimization technique, which has the ability to optimize solutions using the global space of search solutions. Another advantage is the use of the Naive Bayes algorithm as the basic algorithm for Rotation Forest, which gave a slightly smaller increase in efficiency than the PSO optimization.

Fig. 6 below shows all the results obtained during the study. As we can see, the combined method of vectorization of PV-DM and TF-IDF actually shows better results than standard methods. However, it shows a slight improvement over the TF-IDF method.

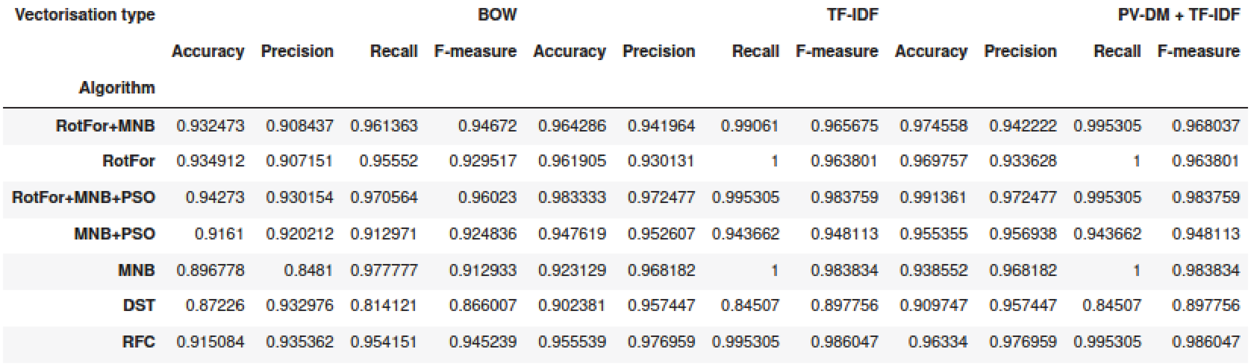
Conclusions

This study was driven by an increase in email spam worldwide. From the literature of spam classification algorithms, it was concluded that in each of the stages of the spam classifier many experiments are conducted and there is a need to create a better classifier that will contain the best of the components. The Rotation Forest method was chosen for improvement, experiments with which showed that the use of another basic algorithm can significantly improve the efficiency of this metamodel.

The experiments were performed on the basis of a dataset, which is a combination of Enron, Ling and SpamAssasin datasets. They showed that the modified Rotation Forest algorithm works with an accuracy of 99.14%, which is 2.17% better than the basic algorithm Rotation Forest.

Making a conclusion from the results of classifiers that worked on data vectorized in different ways, we can say that the particle swarm optimization gives a greater increase in efficiency than the replacement of the basic method by Naive Bayes. But both of these modifications make a significant contribution to the resulting algorithm.

To improve the stability of the classifier, it is recommended to use additionally other datasets, such as Trec 2005, Trec 2006, Trec 2007, Spam archive, etc.



Vectorisation type	BOW				TF-IDF				PV-DM + TF-IDF			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
Algorithm												
RotFor+MNB	0.932473	0.908437	0.961363	0.94672	0.964286	0.941964	0.99061	0.965675	0.974558	0.942222	0.995305	0.968037
RotFor	0.934912	0.907151	0.95552	0.929517	0.961905	0.930131	1	0.963801	0.969757	0.933628	1	0.963801
RotFor+MNB+PSO	0.94273	0.930154	0.970564	0.96023	0.983333	0.972477	0.995305	0.983759	0.991361	0.972477	0.995305	0.983759
MNB+PSO	0.9161	0.920212	0.912971	0.924836	0.947619	0.952607	0.943662	0.948113	0.955355	0.956938	0.943662	0.948113
MNB	0.896778	0.8481	0.977777	0.912933	0.923129	0.968182	1	0.983834	0.938552	0.968182	1	0.983834
DST	0.87226	0.932976	0.814121	0.866007	0.902381	0.957447	0.84507	0.897756	0.909747	0.957447	0.84507	0.897756
RFC	0.915084	0.935362	0.954151	0.945239	0.955539	0.976959	0.995305	0.986047	0.96334	0.976959	0.995305	0.986047

Fig. 6. The results of algorithms, using different methods of vectorization

Because the more diverse the data, the more stable the algorithm. Also, for a qualitative classification of spam regardless of language, you should use datasets in other languages. It is recommended to investigate the spam features within different language groups.

To improve the resulting efficiency of algorithm, it is recommended to try to use PSO-BO optimization, which improves the search for hyperparameters of the model.

For better recognition, you should delve deeper into the most modern methods of vectorization and text preprocessing. Find more appropriate sets of stop-words in spam emails. It may also be advisable not to delete some data, such as punctuation marks or numbers, but to use it in a special way. Also recommended try to use different machine learning tools.

References

1. Hossein Siadati, Sima (Tahereh) Jafarikhah and Markus Jakobsson (2016), Traditional Countermeasures to Unwanted Emails, *Understanding Social Engineering Based Scams*, pp. 51-62. http://dx.doi.org/10.1007/978-1-4939-6457-4_5.
2. The Radicati Group (2017-2021), *Email Statistics Report.*
3. Guzella, T.S. and Caminhas, W.M. (2009), A review of machine learning approaches to Spam filtering, *Expert Systems with Application,* Vol. 36, No. 7, pp. 10206-10222. http://dx.doi.org/10.1016/j.eswa.2009.02.037.
4. Shafi'i Muhammad Abdulhamid, Maryam Shuaib and Oluwafemi Osho (2018), Comparative Analysis of Classification Algorithms for Email Spam Detection, *International Journal of Computer Network and Information Security*, Vol. 1, pp. 60-67. http://dx.doi.org/10.5815/ijcnis.2018.01.07.
5. Simranjit Kaur Tuteja and Bogiri Nagaraju (2016), Email Spam filtering using BPNN classification algorithm, *International Conference on Automatic Control and Dynamic Optimization Techniques*, pp. 915-919. http://dx.doi.org/10.1109/ICACDOT.2016.7877720.
6. Kumaresan, T. and Palanisamy, C. (2017), Email spam classification using S-cuckoo search and support vector machine, *International Journal of Bio-Inspired Computation*, Vol. 9, No. 3, pp. 142-156. https://dx.doi.org/10.1504/IJBIC.2017.083677.
7. Aakanksha Sharaff, Naresh Nagwani and Abhishek Dhadse (2016), Comparative Study of Classification Algorithms for Spam Email Detection, *Emerging Research in Computing, Information, Communication and Applications*, pp. 237-244. http://dx.doi.org/10.1007/978-81-322-2553-9_23.
8. Muhammad Iqbal, Malik Muneeb Abid, Mushtaq Ahmad and Faisal Khurshid (2016), Study on the Effectiveness of Spam Detection Technologies, *International Journal of Information Technology and Computer Science*, Vol. 8, No. 1, pp. 11-21. http://dx.doi.org/10.5815/ijitcs.2016.01.02.
9. Mohammad Zavvar, Meysam Rezaei and Shole Garavand (2017), Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine, *International Journal of Modern Education and Computer Science*, Vol. 8, No. 7, pp. 68-74. http://dx.doi.org/10.5815/ijmecs.2016.07.08.
10. Vishal Kumar Singh and Shweta Bhardwaj (2018), Spam Mail Detection Using Classification Techniques and Global Training Set, *Intelligent Computing and Information and Communication*, pp. 623-632. http:/dx.doi.org/10.1007/978-981-10-7245-1_61.
11. Reena Sharma and Gurjot Kaur (2016), Email Spam Detection Using SVM and RBF, *International Journal of Modern Education and Computer Science*, Vol. 8, No. 4, pp. 57-63. http://dx.doi.org/10.5815/ijmecs.2016.04.07.
12. Rodríguez, J.J., Kuncheva Ludmila and Alonso, C.J. (2006), Rotation Forest: A New Classifier Ensemble Method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 10, pp. 1619-1630. http://dx.doi.org/10.1109/TPAMI.2006.211.
13. Binh Thai Pham, Dieu Tien Bui, Dholakia M.B., Indra Prakash, Ha Viet Pham, Khalid Mehmood and Hung Quoc Le (2015), A Novel Ensemble Classifier of Rotation Forest and Naïve Bayer for Landslide Susceptibility Assessment at the Luc Yen District, Yen Bai Province (Viet Nam) Using GIS, *Geomatics, Natural Hazards and Risk*, pp. 649-671. https://doi.org/10.1080/19475705.2016.1255667.
14. Borja Ayerdi and Manuel Graña (2016), Anticipative Hybrid Extreme Rotation Forest, *Procedia Computer Science*, Vol. 80, pp. 1671-1681. https://doi.org/10.1016/j.procs.2016.05.507.
15. Nandan Parmar, Ankita Sharma, Harshita Jain and Dr. Kadam, A.K. (2018), Email Spam Detection using Naïve Bayes and Particle Swarm Optimization, *Second International Conference on Intelligent Computing and Control Systems*, pp. 685-690. https://doi.org/10.1109/ICCONS.2018.8662957.
16. Breaban, Mihaela, Ionita, Madalina and Croitoru, Cornelius (2007), A new PSO approach to constraint satisfaction, *IEEE Congress on Evolutionary Computation,* pp. 1948-1954. https://doi.org/10.1109/CEC.2007.4424712.
17. Sutta, N., Liu, Z. and Zhang, X. (2020), A Study of Machine Learning Algorithms on Email Spam Classification, *35th International Conference on Computers and Their Applications*, Vol. 69, pp. 170-179. https://doi.org/10.29007/qshd.
18. Bozkir, A.S, Esra Sahin, Murat Aydos and Ebru Akcapinar Sezer (2017), Spam Email Classification by Utilizing N-Gram Features of Hyperlink Texts, *IEEE 11th International Conference on Application of Information and Communication Technologies*, pp. 1-5. https://doi.org/10.1109/ICAICT.2017.8687020.
19. Samira Douzi, AlShahwan, F.A., Mouad Lemoudden and Bouabid El Ouahidi (2020), Hybrid Email Spam Detection Model Using Artificial Intelligence, *International Journal of Machine Learning and Computing*, Vol. 10, No. 2, pp. 316-322. http://dx.doi.org/10.18178/ijmlc.2020.10.2.937.
20. Mikolov, Tomas, Sutskever, Ilya, Kai Chen, Corrado, G.S. and Jeffrey Dean (2013), Distributed Representations of Words and Phrases and their Compositionality, *Neural Information Processing Systems*, pp. 3111-3119. Available at: https://arxiv.org/abs/1310.4546.
21. Charbonnier, Jean and Wartena, Christian (2018), Using Word Embeddings for Unsupervised Acronym Disambiguation, *The 27th International Conference on Computational Linguistics*, pp. 2610-2619.

Список літератури

1. Hossein Siadati, Sima (Tahereh) Jafarikhah, Markus Jakobsson. Traditional Countermeasures to Unwanted Emails. *Understanding Social Engineering Based Scams*. 2016. P. 51-62. http://dx.doi.org/10.1007/978-1-4939-6457-4_5.

2. The Radicati Group. Email Statistics Report, 2017-2021.

3. Guzella Thiago S., Caminhas W.M. A review of machine learning approaches to Spam filtering. *Expert Systems with Applications.* 2009. Vol. 36. No. 7. P. 10206-10222. http://dx.doi.org/10.1016/j.eswa.2009.02.037.

4. Shafi'i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho. Comparative Analysis of Classification Algorithms for Email Spam Detection. *International Journal of Computer Network and Information Security*. 2018. Vol. 1. P. 60-67. http://dx.doi.org/10.5815/ijcnis.2018.01.07.

5. Simranjit Kaur Tuteja, Bogiri Nagaraju. Email Spam filtering using BPNN classification algorithm. *International Conference on Automatic Control and Dynamic Optimization Techniques*. 2016. P. 915-919. http://dx.doi.org/10.1109/ICACDOT.2016.7877720.

6. Kumaresan T., PalanisamyC. Email spam classification using S-cuckoo search and support vector machine. *International Journal of Bio-Inspired Computation.* 2017. Vol. 9. No. 3. P. 142-156. https://dx.doi.org/10.1504/IJBIC.2017.083677.

7. Aakanksha Sharaff, Naresh Nagwani, Abhishek Dhadse. Comparative Study of Classification Algorithms for Spam Email Detection. *Emerging Research in Computing, Information, Communication and Applications*. 2016. P. 237-244. http://dx.doi.org/10.1007/978-81-322-2553-9_23.

8. Muhammad Iqbal, Malik Muneeb Abid, Mushtaq Ahmad, Faisal Khurshid. Study on the Effectiveness of Spam Detection Technologies. *International Journal of Information Technology and Computer Science*. 2016. Vol. 8, No. 1, P. 11–21. http://dx.doi.org/10.5815/ijitcs.2016.01.02.

9. Mohammad Zavvar, Meysam Rezaei, Shole Garavand. Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine. *International Journal of Modern Education and Computer Science*. 2016. Vol. 8, No. 7. P. 68-74. http://dx.doi.org/10.5815/ijmecs.2016.07.08.

10. Vishal Kumar Singh, Shweta Bhardwaj. Spam Mail Detection Using Classification Techniques and Global Training Set. *Intelligent Computing and Information and Communication*. 2018. P. 623-632. http:/dx.doi.org/10.1007/978-981-10-7245-1_61.

11. Reena Sharma, Gurjot Kaur. Email Spam Detection Using SVM and RBF. *International Journal of Modern Education and Computer Science*. 2016. Vol. 8. No. 4. P. 57-63. http://dx.doi.org/10.5815/ijmecs.2016.04.07.

12. Rodríguez Juan J., Kuncheva Ludmila, Alonso Carlos J. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. October 2006. Vol. 28. No. 10. P. 1619-1630. http://dx.doi.org/10.1109/TPAMI.2006.211.

13. Binh Thai Pham, Dieu Tien Bui, Dholakia M.B., Indra Prakash, Ha Viet Pham, Khalid Mehmood, Hung Quoc Le. A Novel Ensemble Classifier of Rotation Forest and Naïve Bayer for Landslide Susceptibility Assessment at the Luc Yen District, Yen Bai Province (Viet Nam) Using GIS. *Geomatics, Natural Hazards and Risk*. P. 649-671. https://doi.org/10.1080/19475705.2016.1255667.

14. Borja Ayerdi, Manuel Graña. Anticipative Hybrid Extreme Rotation Forest. *Procedia Computer Science*. 2016. Vol. 80, P. 1671-1681. https://doi.org/10.1016/j.procs.2016.05.507.

15. Nandan Parmar, Ankita Sharma, Harshita Jain, Dr. Amol K. Kadam. Email Spam Detection using Naïve Bayes and Particle Swarm Optimization. *Second International Conference on Intelligent Computing and Control Systems*. 2018. P. 685-690. https://doi.org/10.1109/ICCONS.2018.8662957.

16. Breaban Mihaela, Ionita Madalina, Croitoru Cornelius. A new PSO approach to constraint satisfaction. *IEEE Congress on Evolutionary Computation*. 2007. P. 1948-1954. https://doi.org/10.1109/CEC.2007.4424712.

17. N. Sutta, Liu Z., Zhang X. A Study of Machine Learning Algorithms on Email Spam Classification. *35th International Conference on Computers and Their Applications*. 2020. Vol. 69. P. 170-179. https://doi.org/10.29007/qshd.

18. Bozkir Selman, Esra Sahin, Murat Aydos, Ebru Akcapinar Sezer. Spam Email Classification by Utilizing N-Gram Features of Hyperlink Texts. *IEEE 11th International Conference on Application of Information and Communication Technologies*. 2017. P. 1-5. https://doi.org/10.1109/ICAICT.2017.8687020.

19. Samira Douzi, Feda A. AlShahwan, Mouad Lemoudden, Bouabid El Ouahidi. Hybrid Email Spam Detection Model Using Artificial Intelligence. *International Journal of Machine Learning and Computing*. 2020. Vol. 10. No. 2. P. 316-322. http://dx.doi.org/10.18178/ijmlc.2020.10.2.937.

20. Mikolov Tomas, Sutskever Ilya, Kai Chen, Corrado G.S., Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems*. 2013. P. 3111-3119. Available at: https://arxiv.org/abs/1310.4546.

21. Charbonnier Jean, Wartena Christian. Using Word Embeddings for Unsupervised Acronym Disambiguation. *The 27th International Conference on Computational Linguistics*. 2018. P. 2610-2619.

Received by Editorial Board 10.02.2021
Signed for Printing 02.03.2021

Відомості про автора:

Шанін Андрій Олександрович
магістрант
студент Харківського національного
університету радіоелектроніки,
Харків, Україна
https://orcid.org/0000-0002-0043-4613

Information about the author:

Andrii Shanin
Graduate Student
of Kharkiv National University
of Radio Electronics,
Kharkiv, Ukraine
https://orcid.org/0000-0002-0043-4613

МОДИФІКАЦИЯ МОДЕЛІ ROTATION FOREST В РАМКАХ ЗАДАЧІ КЛАСИФІКАЦІЇ СПАМУ ЕЛЕКТРОННОЇ ПОШТИ

А.О. Шанін

Збільшення використання електронної пошти в щоденних транзакціях для багатьох підприємств або загального спілкування завдяки своїй економічній ефективності зробило електронні листи вразливими до атак, включаючи спам. Спам-листи – це небажані повідомлення, які дуже схожі один до одного та надсилаються декільком одержувачам випадковим чином. Аналізуючи останні дослідження та публікації в цій галузі, було зроблено висновок, що найбільш якісним способом векторизації тексту для подальшої класифікації є поєднання методів PV-DM та TF-IDF, а найкраща модель для класифікації спаму це Rotation Forest. Отже, метою цього дослідження є модифікація моделі Rotation Forest та створення найбільш якісного класифікатора для задачі класифікації спаму електронної пошти. Оскільки алгоритм Naive Bayes в рамках класифікації спаму працює набагато краще, ніж Decision Tree, було вирішено використовувати алгоритм Naive Bayes як базовий алгоритм у модифікованій моделі Rotation Forest. Виходячи з результатів досліджень методів оптимізацій, виявилось що оптимізація рою частинок (PSO) значно покращує ефективність алгоритму Naive Bayes в рамках класифікації спаму. Тому для тренування базових слабких алгоритмів також застосовували оптимізацію PSO. Для поліпшення стабільності класифікатора експерименти проводились на основі комбінації Enron, Ling та SpamAssasin датасетів і оцінювались з точки зору точності (accuracy), f-міри (f-measure), влучності (precision) та повноти (recall). В результаті експериментів було показано, що запропонований модифікований алгоритм Rotation Forest дійсно працює значно краще відносно стандартного алгоритму Rotation Forest. Модифікований алгоритм Rotation Forest показав високу точність класифікації в 99,14%, тоді як стандартний Rotation Forest працює з точністю 96,97%. В результаті дослідження ми створили справді якісний класифікатор. Однак, оскільки точність класифікації не є 100%, цьому алгоритму є куди рости.

Keywords: *Email spam classification, Rotation Forest, Naive Bayes, Particle Swarm Optimisation, PV-DM, TF-IDF.*

МОДИФИКАЦИЯ МОДЕЛИ ROTATION FOREST В РАМКАХ ЗАДАЧИ КЛАССИФИКАЦИИ СПАМА ЭЛЕКТРОННОЙ ПОЧТЫ

А.А. Шанин

Увеличение использования электронной почты в ежедневных транзакциях для многих предприятий или общего общения благодаря своей экономической эффективности сделало электронные письма уязвимыми к атакам, включая спам. Спам-письма – это нежелательные сообщения, которые очень похожи друг на друга и направляются нескольким получателям случайным образом. Анализируя последние исследования и публикации в этой области, был сделан вывод, что наиболее качественным способом векторизации текста для дальнейшей классификации является сочетание методов PV-DM и TF-IDF, а лучшая модель для классификации спама это Rotation Forest. Итак, целью данного исследования является модификация модели Rotation Forest и создание наиболее качественного классификатора для задачи классификации спама электронной почты. Поскольку алгоритм Naive Bayes в рамках классификации спама работает гораздо лучше, чем Decision Tree, было решено использовать алгоритм Naive Bayes как базовый алгоритм в модифицированной модели Rotation Forest. Исходя из результатов исследований методов оптимизаций, оказалось, что оптимизация роя частиц (PSO) значительно улучшает эффективность алгоритма Naive Bayes в рамках классификации спама. Поэтому для тренировки базовых слабых алгоритмов также применяли оптимизацию PSO. Для улучшения стабильности классификатора эксперименты проводились на основе комбинации Enron, Ling и SpamAssasin датасетов и оценивались с точки зрения точности (accuracy), f-меры (f-measure), меткости (precision) и полноты (recall). В результате экспериментов было показано, что предложенный модифицированный алгоритм Rotation Forest действительно работает значительно лучше относительно стандартного алгоритма Rotation Forest. Модифицированный алгоритм Rotation Forest показал высокую точность классификации в 99,14%, тогда как стандартный Rotation Forest работает с точностью 96,97%. В результате исследования мы создали действительно качественный классификатор. Однако, поскольку точность классификации не является 100%, этому алгоритму есть куда расти.

Keywords: *Email spam classification, Rotation Forest, Nauve Bayes, Particle Swarm Optimisation, PV-DM, TF-IDF.*