DOI: 10.21917/ijsc.2020.0316

K-NEAREST NEIGHBOR CLASSIFICATION OF E-MAIL MESSAGES FOR SPAM DETECTION

U. Murugavel and R. Santhi
Department of Computer Science, Bharathiar University, India

Abstract

The increase in unwanted spam email volumes created a clear need for more effective and robust anti-spam filters. Recent machine learning methods are employed to detect and process spam mails successfully. In this paper, we present a density based clustering for email classification problem using kNN algorithm. Initially, the relevant features for filtering the spam messages are extracted from the study and it acts as an antispam filter. It thereby generates the successful corpus list for detection of spam emails. The experiments are conducted on various email datasets and the results show that the proposed kNN density based clustering offers improved performance than the other methods. As shown by the test results, our methodology showed stronger prediction capabilities and better classifications based on in-depth learning techniques.

Keywords:

Email, Spam Classification, Feature Extraction, Corpus

1. INTRODUCTION

As Internet users are developing rapidly, email is a key tool and an easier way to transmit data and share useful information about users to communicate in an electronic medium within a minute. Millions of users use e-mail every day to share information both personally and financially. In case of this, there will be more option for unwanted, unlawful spam emails as junk mails and occupy additional stock and the annoying task of email users will be to handle them. Ethical hacking is a robust mechanism to identify spammers' unsolicited e-mail messages and network weaknesses. The ethical hacking techniques handling spam thread countermeasures receive millions of spammers' mails uninvited. The ethical hacking of spam threads is performed legally by well trained professionals. Ethical hacker is also called white hat or penetration testing system, which can control spam threads in a legally binding way.

Spam mails can be either unwanted bulk emails or unwanted business emails. Spam mails contain promising different offerings for the user by means of electronic messages, junk e-mails spam overflows the inbox which annoy the users to use it, malware spam includes malware emails with virus, or spam spam mail is a problem for the user and user of spam mail is an important element.

Spammers can sometimes use a new way of delivering irrelevant messages in the form of Image Spam that breaks the clear-cut method most of which use spam images etc. Text analysis is used to extract the most common spam words in the user's inbox using content based analytics. Spam is an irrelevant message, which is generally sent by internet users to a large number of users for email spoofing, cash scams, cords, hoaxes, ads and malware on the internet. Efficient and rigid spam filtering mechanism is needed to deal with the extensive spam e-mails. In this paper various ways of using text-analytics methods to decrease spam were discussed by extracting frequent spam words, and processing them efficiently and most often to identify the spam thread to solve the spam problem.

In the work currently under way, it was only determined whether the inward email is spam or ham. The current work has an issue in which the spam threads are categorised under which spam threads are set. The proposed work will examine the spam words from the spam corpus database and will bring spam threads together by checking and analysing spam with the spam keywords that match the spam and identifying the email users' most influential spam threads.

In this article we present an email classification problem clustering based on the density by means of kNN algorithm. Initially the relevant functionality is extracted from the trial and acts like an anti-spam filter for filtering the spam messages. The successful corpus list for spam emails is therefore produced. The experiments are performed on different email data sets and the results demonstrate that the proposed clustering based on kNN density provides better performance than the other methods. Our methodology demonstrated stronger prediction capacity and better classifications based on thorough learning techniques, as demonstrated by the results.

2. RELATED WORKS

The authors in [1] presented the different prevalent methods of spam filtering and recognised the drawback on the filtering method based on content. The paper focused on the existing work and identified the wrong words on the Bayesian filter based on content. The work extracts more precision than normal spam filters.

Different decision tree classifiers were applied in [2] to separate spam and ham mail. Different filtering methods were used to filter the spam mails. The work analysed weka tool results and compared different classifiers' experimental results.

The authors presented in [3] that spam mails need to be detected. The paper focused on different techniques for spam detection to reduce spam.

The authors of [4] have been able to categorise spam mail using the Tanagra tool. The paper also extracts the attribute by selecting attributes. Data were applied to different classifier and the results were analysed with cross validation methods. The test results show spam on the basis of error, accuracy and retrieval parameters.

The authors of [5] focused on spam and zombie attack prediction. The paper developed a tool for predicting spam using an algorithm for spot detection.

The authors described in [6] the existing spam filtering methods. The paper used methods of evaluation based on learning. The paper compared the anti-spam data and examined some new spam filtering techniques [6].

The authors discussed in [7] the assessment of hacking, which explores the links and efficiently demonstrates ethical hacking.

The authors in [8] presented the various ways of finding a social network spammer. This paper examined the spam detection social network Twitter. In this work, the spam is categorised according to false content, spam URL, spam and trend topics. Different functions include user, content, graph, structure and spam detection time.

In [9] the authors concentrated on e-mail spam detection machine learning methods. The article contains the different components of the email structures including headers, SMTP envelope, data header and email body. The work focused on smart spam detection in email and discussed the different techniques efficiently and effectively.

In [10], the authors classified electronic mail as legally permissible mail and spam. In this paper, efficient spam filters separating the spam mail were required, including the paper using either the basis of content or the header base [10].

In [11] the authors concentrated on social network Twitter which classified spam tweets based on content. The paper tweeted the text and used filtering compare algorithms.

In [12] the authors imposed spam-free content characteristics on spam detection. The paper summarised the pages and outlined the entropy and independent n-grams measures for improved results. The paper also made calculations based on the correlation of multiple features.

In [13], authors presented previously undescribed techniques which automatically detect the spam document and examine classification algorithms based on efficiency. The paper has more efficiently identified spam or ham messages.

In [14] the authors formulated a prototype for the classification of spam images. The paper extracts an image-based image detection classifier. The paper achieves optimised spam discovery accuracy.

3. PROPOSED METHOD

The proposed approach is used to determine the implicit information and explicit information and a mechanism to validate its ability to extract the messages. The major process is involves: pre-processing process, clustering process and classification process. The data-preprocessing involves certain modules that involve cleaning of input data, selection of attributes, transformation of data and integration of data. Finally, K-NN classification algorithm after feature selection is used to classify the datasets, which are used to classify the selected attributes for crop yield, which sets the condition that the crops yield is possible or not.

3.1 DATA PREPROCESSING

The quality of datasets are improved using initial pre-processing operation that sets the input data for further clustering process i.e. it makes the data fit for clustering process. The high-quality data provide high output quality and this gives the information or knowledge to predict crop yield. The four major steps in pre-processing step involves the following operations.

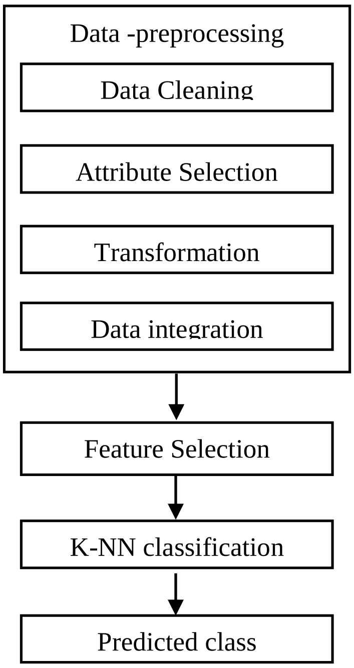

Fig.1. Proposed framework using KNN classification

3.2 DATA CLEANING

The process of data cleaning removes from the missing or incomplete or noisy or inconsistent data. Data purification ensures that the quality of the knowledge acquired is high. The cleaning of data is profoundly domain-specific. Problems in data quality are fairly trivial, complex and uniform. There is no common international reference standard. The process is therefore different between the domain and the domain, but is essentially used to determine inaccurate, incomplete or unreasonable data and then to improve quality by correcting omissions and detected errors. This usually leads to flagging, documentation and subsequent checking and suspected records correction. Checking validation may also involve verification that standards, rules and conventions are complied with. Data cleaning principle is to identify and correct discrepancies and errors.

3.2.1 Attribute Selection:

The selection of the correct attributes contributes to better data extraction. Removal of above- mentioned redundant information enhances the selection of attributes. The data space eliminates low information gain attributes, involving poor seed quality and crop spacing and irrigation method.

3.2.2 Transformation:

The process of transformation covers the data into a better form used for the mining process. The numerical or integer or nominal datasets are converted to categorical datasets.

3.2.3 Data Integration:

The data for the integration process are gathered from farmers from various regions. For the clustering process, the collected data is integrated further. For use in the clustering process, the data must be ready and passed through several steps, as described below:

Step 1. Convert text data format into categorical data format.

Step 2. Divide the features in the categorical data format into three categories: Season, Crop and Area.

Step 3. Estimate the association or correlation between the features of datasets i.e. predictors and attributes, and the response attributes.

Step 4. Find and select the relevant features w.r.t response attributes.

Step 5. The total size of clusters is reduced using clustering process that eliminates huge attributes or features from the relevant features.

3.3 KNN CLASSIFICATION

K nearest neighbors is considered as a simple algorithm which stores and classifies all available cases based on similarity measures. The classification method defines the class as yield or not-yield. A stable and efficient classification method based on examples is the KNN algorithm. The process of classifying a document is as follows using the KNN algorithm:

In the document set, the similar *K* training documents for a particular trial documentation *d* are found. Then each document class has a value that represents the similitude amount between the test documentation and the class *K* training documentation. In the *K* documents, that is, if there are class document, the amount of similarity between those documents and the test documentation is the value of this class. When selecting scores, we take the K documents into consideration only the score more than the threshold after the statistical value of the class.

The steps of k-NN classification algorithm are given below:

Step 1. Assume the nearest number (*K*) from a class;

Step 2. Find the similarly measurement between the entire training sets and the test documents *d*.

Step 3. Select *K* documents that lie to be more similar w.r.t test documents *d* as its closest document *d*.

Step 4. Collect the similar classes from the neighbor documents

Step 5. Provide a value to a class using nearest *K* documents

$$Score(d,c_i) = \sum_{d_j \in KNN} Sim(d,d_j)\, y(d_j,c_i) - b_i$$

$$for\ y(d_j,c_i) = \begin{cases} 1 & d_j \in c_i \\ 0 & d_j \notin c_i \end{cases}, b_i \text{ is threshold}$$

Select the biggest value in a class and consider it as an appropriate test document from the test document.

4. PERFORMANCE EVALUATION

This is a CSV file containing related information from 5172 randomly selected email files and their labels. Each row for each email is 5172 rows in the csv file. 3002 columns are available. Email Name is shown in the first column. The name is numbered and not the name of the recipient for protection of privacy. The labelling for prediction is given in the last column: 1 for spam, 0 for spam. There are still three thousand columns in each email, after the non-alphabetically character/words are excluded.

Table.1 shows the comparison of Execution Time between the KNN and existing classifiers for various spam messages. Table.2 shows the comparison of Computational Overhead between the KNN and existing classifiers for various spam messages. Table.3 shows the comparison of Detection rate between the KNN and existing classifiers for various spam messages. Table.4 shows the comparison of Security impact between the KNN and existing classifiers for various spam messages. Table.5 shows the comparison of Energy consumption between the KNN and existing classifiers for various spam messages. Table.6 shows the comparison of Memory requirement between the KNN and existing classifiers for various spam messages.

Table.1. Execution Time

Classifiers	Execution Time
K-Nearest Neighbor	0.96138
Random Forest	0.96293
K-means	0.963646
Decision Tree	0.964219
Naïve Bayes	0.965951
Perceptron Rule Base	0.966107

Table.2. Detection rate

Classifiers	Detection rate
K-Nearest Neighbor	0.604428
Random Forest	0.622364
K-means	0.625737
Decision Tree	0.630347
Naïve Bayes	0.645609
Perceptron Rule Base	0.647051

Table.3. Memory requirement

Classifiers	Memory requirement
K-Nearest Neighbor	0.32115
Random Forest	0.313834
K-means	0.304943
Decision Tree	0.301431
Naïve Bayes	0.283194
Perceptron Rule Base	0.274924

Table.4. Computational Overhead

Classifiers	Computational Overhead
K-Nearest Neighbor	0.544712
Random Forest	0.568991
K-means	0.569417
Decision Tree	0.574265
Naïve Bayes	0.58184
Perceptron Rule Base	0.589668

Table.5. Security impact

Classifiers	Security impact
K-Nearest Neighbor	0.67885
Random Forest	0.686166
K-means	0.695057
Decision Tree	0.698569
Naïve Bayes	0.716806
Perceptron Rule Base	0.725076

Table.6. Average Energy consumption Rate

Classifiers	Energy consumption
K-Nearest Neighbor	0.974218
Random Forest	0.975825
K-means	0.975927
Decision Tree	0.97635
Naïve Bayes	0.976826
Perceptron Rule Base	0.977293

5. CONCLUSION

In this chapter, we propose a framework to improve the prediction by the selection of correct attributes that contributes to better data extraction. Removal of redundant information enhances the selection of attributes. This series of frameworks achieves the aim of improving clustering quality. Predictive quality is increased by classification based on the K-NN classification improves the ability of classifying the datasets than other methods.

REFERENCES

[1] D. Gaurav, S.M. Tiwari, A. Goyal, N. Gandhi and A. Abraham, “Machine Intelligence-Based Algorithms for Spam Filtering on Document Labeling”, *Soft Computing*, Vol. 24, No. 13, pp. 9625-9638, 2020.

[2] A. Bhowmick and S.M. Hazarika, “E-Mail Spam Filtering: A Review of Techniques and Trends”, *Proceedings of International Conference on Advances in Electronics, Communication and Computing*, pp. 583-590, 2018.

[3] M.K. Chae, A. Alsadoon, P.W.C. Prasad and S. Sreedharan, “Spam Filtering Email Classification (SFECM) using Gain and Graph Mining Algorithm”, *Proceedings of International Conference on Anti-Cyber Crimes*, pp. 217-222, 2017.

[4] A.S. Aski and N.K. Sourati, “Proposed Efficient Algorithm to Filter Spam using Machine Learning Techniques”, *Pacific Science Review A: Natural Science and Engineering*, Vol. 18, No. 2, pp. 145-149, 2016.

[5] P.M. Paul and R. Ravi, “A Collaborative Reputation-Based Vector Space Model for Email Spam Filtering”, *Journal of Computational and Theoretical Nanoscience*, Vol. 15, No. 2, pp. 474-479, 2018.

[6] B.K. Dedeturk and B. Akay, “Spam Filtering using a Logistic Regression Model Trained by an Artificial Bee Colony Algorithm”, *Applied Soft Computing*, Vol. 91, pp. 1-17, 2020.

[7] M. Sharma and S. Sharma, “A Survey of Email Spam Filtering Methods”, *Control Theory and Informatics*, Vol. 7, pp. 2224-5774, 2018.

[8] J.R. Mendez, T.R. Cotos Yanez and D. Ruano Ordas, “A New Semantic-Based Feature Selection Method for Spam Filtering”, *Applied Soft Computing*, Vol. 76, pp. 89-104, 2019.

[9] N.K. Nagwani and A. Sharaff, “SMS Spam Filtering and Thread Identification using Bi-Level Text Classification and Clustering Techniques”, *Journal of Information Science*, Vol. 43, No. 1. pp. 75-87, 2017.

[10] A. Barushka and P. Hajek, “Spam Filtering using Integrated Distribution-Based Balancing Approach and Regularized Deep Neural Networks”, *Applied Intelligence*, Vol. 48, No. 10, pp. 3538-3556, 2018.

[11] H. Yang, Q. Liu, S. Zhou and Y. Luo, “A Spam Filtering Method based on Multi-Modal Fusion”, *Applied Sciences*, Vol. 9, No. 6, pp. 1152-1163, 2019.

[12] M. Shuaib, O.S. Adebayo, O. Osho, I. Idris, J.K. Alhassan and N. Rana, “Whale Optimization Algorithm-based Email Spam Feature Selection Method using Rotation Forest Algorithm for Classification”, *SN Applied Sciences*, Vol. 1, No. 5, pp. 390-406, 2019.

[13] T.A. Almeida, T.P. Silva, I. Santos and J.M.G. Hidalgo, J. M. G. (2016). Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering. Knowledge-Based Systems, 108, 25-32, 2016.

[14] M.A.E.S. Mokri, R.M. Hamou and A. Amine, “A New Bio Inspired Technique based on Octopods for Spam Filtering”, *Applied Intelligence*, Vol. 49, No. 9, pp. 3425-3435, 2019.