EMAIL SPAM FILTERING USING DECISION TREE ALGORITHM

Divesh Palival, Kevin Printer, Ramchandra Devre, Asst.Prof. Nikita Lemos

Abstract— E-mails have become the best way to communicate formal documents over the Internet among users. But many people have started sending unwanted mails to others also called as email spam. Many email spam messages are commercial in nature but may also contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware. Spam email sometimes include malware or viruses or any other other executable file attachments which are very risky and can cause serious threats to one's system. Several spam filtering systems exist in real world to filter spam mails like Blacklisting, Signature based System. In Blacklisting, received mail is checked by a mail server IP address against a pool of email blacklist. So if anyone's mail server has been blacklisted then his/her mail will not be forwarded. The disadvantage in this system is that it leads to high false negative rate which makes them unreliable. Signature based system compares any incoming mail to a known spam by computing its signature. But it is very inefficient since it catches only 60-70% of spam mails. To handle all the above mentioned limitations we are using ID3 algorithm. The ID3 algorithm is based on the Decision tree algorithm. ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations (in our case, Enron dataset). The resulting tree is used to classify test observations. Each observation is represented by features or attributes and a class to which it belongs. ID3 uses information gain measure to select decision node. Information gain indicates the ability of a given attribute to separate training examples into classes. Higher the information gain, higher is the ability of the attribute to separate training observation. Information gain uses entropy as a measure to calculate the amount of uncertainty in dataset.

Index Terms— spam, ham, ID3 algorithm, email spam detection, Information gain , Enron

1 INTRODUCTION

Email allows a user to send and receive messages to and from anyone with an email address, anywhere in the world without spending a penny. It can be accessed from anywhere in the world and can deliver messages instantaneously. Because the mobile access to email is not attached to a physical location, the mobility of email allows people to work and communicate from anywhere. Due to these factors, email communication is used over other modes of communication because it is economical, flexible and reasonable. Such a benefit of a communication is sometimes misused and overused which in turn leads to a growth in spam mails. A spam is a type of electronic spam where irrelevant or unrequested messages are sent by email. Malwares as scripts or other executable file attachments which may be viruses may also be included in the spam email. To segregate between the legitimate and the spam or junk mails makes it very tedious for the user. A lot of unwanted space in the memory is hogged up by the spam mails. These spam mails can be removed manually which makes it very undesirable and inconvenient to the user.. These spam mails make it very tedious for the user to segregate between the legitimate (any mail which is acceptable or recognized as genuine or valid) and the spam or junk mails. Such spam mails consume a lot of memory which hogs up unwanted space. These spam mails can be removed manually which makes it very undesirable and inconvenient to the user.

The statistics related to spam are described in the Fig.No.1.1. The table shows the spam during the month of January, 2018. It also showcases the percentage of email accounted as spam as well as the average amount of spam emails generated every second. Hence there is an immense need to avoid the spam mails.

Email accounted as spam	85.27% of all mail
Daily spam mails sent globally	421.81 billion
Average spam	3.9 per second
Maximum spam	12 per second
Spam complaints and reports	85,499

Fig.No.1.1

To filter these spam mails several spam filtering systems exist in real world to filter spam mails like Blacklisting, Signature based System. In Blacklisting, received mail is checked by a mail server IP address against a list of email blacklist. So if anyone's mail server has been blacklisted then his/her mail will not be forwarded. The disadvantage in this system is that it leads to high false negative rate which makes them unreliable. Signature based system compares any incoming mail to a known spam by computing its signature. But it is very inefficient since it catches only 60-70% of spam mails.

Because these methods were not reliable to use, text based techniques were introduced in order to get efficient, consistent and accurate results. To handle all the above mentioned limitations we are using ID3 algorithm. The ID3 algorithm is based on the Decision tree algorithm. ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations. The resulting tree is used to classify test observations. Each observation is represented by features or attributes and a class to which it belongs. ID3 uses information gain measure to select decision node. Information gain indicates the ability

of a given attribute to separate training examples into classes.

2 Related Work

In their paper [1] the authors, A. K. Sharma and S. Sahni have performed a comparative study of the classification algorithms for spam email data analysis by conducting an experiment in the Weka environment by using four algorithms namely ID3, J48, Simple CART and Alternating Decision tree on the spam email data set. These algorithms were later compared on the basis of their classification accuracy. Out of 4601 email instances, ADTree and SimpleCART incorrectly classified 418 and 339 mails respectively which is quite unsatisfactory.

R. K. Kumar, G. Poonkuzhali, and P. Sudhakar[2] performed spam analysis using TANGARA data mining tool to explore the efficient classifier for email spam classification. Relevant features are extracted using the process of feature extraction and selection. Then various classification algorithms are carried out over this dataset and cross validation is achieved for each of those classifiers.Naive Bayes and Multilogical Logistic Regression gave the lowest performance. They had highest error rates with the values 0.1135 and 0.1117 respectively.

In their paper [3] the authors, A. Chharia and R.K. Gupta proposed an elementary classifier combination, diversified both by feature set and different classifiers. The proposed ensemble combines multiple classifiers in four levels with the classifiers being interdependent on the pevious results. Also, the proposed scheme uses meta-learning technique. The final decision is made using the classifiers prediction, their probability of prediction and some combining rules to classify legitimate and spam mails more precisely. The Naive Bayes algorithm performed poorly with the accuracy rate of 86.4%.

A. Iyer, A. Pandey and D. Pamnani had proposed a paper [4] on email filtering and analysis using classification algorithm viz. Naïve Bayes and C4.5 algorithms. They try to provide an inside into these commonly used algorithms, their effectiveness and how classification and data mining approach can simplify the users task and provide a better human interface.

H. Kaur and A. Sharma proposed an improved email spam classification method using integrated particle swarm optimization and Decision tree in their paper [5]. The existing techniques are limited to various significant features of emails utilizing more features resulting in more significant results. They have used integrated particle swarm optimization technique which is based on Decision tree algorithm with unsupervised filtering that enhances the accuracy rate further.

3 System Architecture

The system architecture shown in the Fig.No.3.1 shows the various components of our project. It provides a conceptual model that defines the structure, behavior and more views of our project.

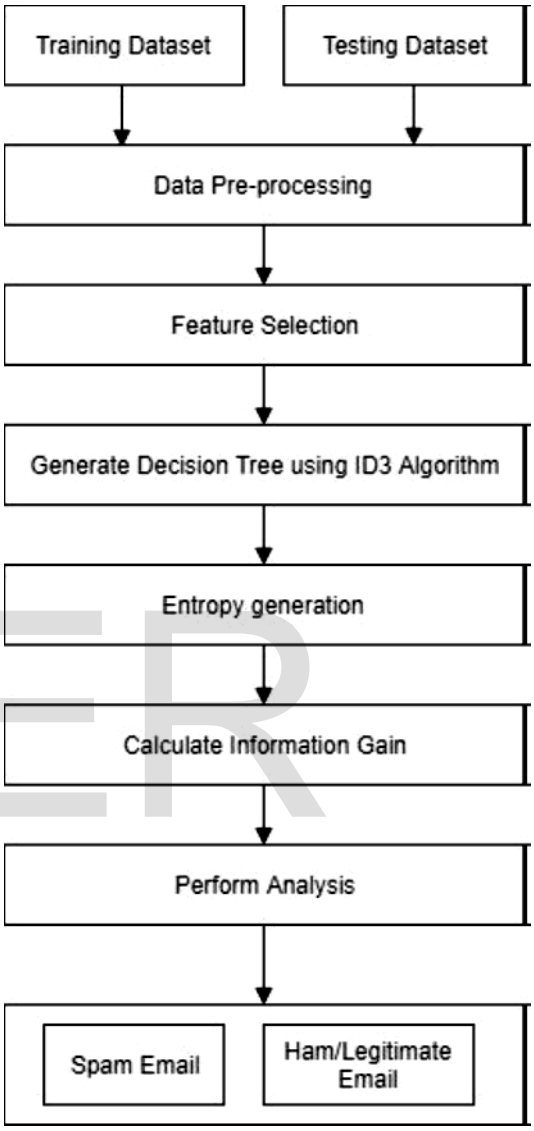

Fig 1: System Architecture

Step 1: Dataset:

The Enron dataset [6] will be used for training as well as testing the filter system. The Enron dataset contains emails of both types stored in plain text format. The Enron directory contains 3672 legitimate (ham) emails and 1500 spam emails. The dataset will be divided into a ratio of 70 : 30 wherein the 70% data will be used for training the system and the remaining 30% will be used for testing the accuracy of the system being developed.

Step 2: Data Pre-processing:

Pre-processing is a very crucial step in text mining. There are three steps involved in pre-processing viz. tokenization, stop word removal and stemming. The initial step consist of the process called tokenization. In this process all the unnecessary word, punctuations and symbols would be removed from the sentences. Now the strings that are left would be split up into various tokens. The next step is stop word removal. Stop-words are the words which carry nearly no information when considered form the text mining point of view. These words contain pronouns, prepositions and conjunctions like he, she, they, and, if, but, etc. In the second step all such words which carry no information are removed. English language has around 300-400 stop words. A list of these words could be made quite easily and referring to that list they can be removed form the sentences, which in turn would save a lot of space required to store them as well as would reduce the operational time to a great extent.

Step 3: Feature Selection:

In this process we analyze the data (emails in our case) minutely to find out the features(i.e. words) which would be most useful in the classification. Then these features would be further used to train the classifier. For this purpose, we will be using the method known as Term Frequency(TF). TF can be defined as a numerical statistic which is intended to reflect how crucial a word is to a document present in a corpus. The TF value is directly proportional to the number of times a word appears in a document.

Step 4: Generate Decision Tree:

The Decision Tree is generated based on the data provided to the system.

Step 5: Entropy generation:

Entropy is defined as a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on c different values, then the entropy S relative to this c-wise classification is defined as

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i c_i$$

where p is the proportion/probability of S belonging to class i. Logarithm to the base 2 is taken because entropy is a measure of the expected encoding length measured in bits.

For e.g. consider a training data having 20 instances with 8 positive and 12 negative instances, the entropy is calculated as

$$Entropy([8+,12-]) = -(\frac{8}{20})\log_2(\frac{8}{20}) - (\frac{12}{20})\log_2(\frac{12}{20}) = 0.9709$$

Step 6: Calculate Information Gain:

Information gain is calculated to split the attributes further in the tree. The attribute with the highest information gain is always preferred first.

Step 7: Performance Analysis:

The performance of the filter will be based on the parameters-True Positive (TP) and True Negative (TN). TP is the condition where the emails provided are spam emails and identified as a spam while TN is the condition where the emails provided are legitimate and identified as legitimate mails.

Step 8: Mark Spam/ Legitimate :

Based on the results obtained from the performance analysis, the legitimacy of the email is determined. A threshold value is set which will be used for comparison between the obtained results and the set threshold. Thus, the email will be classified as spam or legitimate email.

4 Conclusion

The proposed system will detect the spam emails sent and received and will give notification to the respective user. The system uses the ID3 algorithm and decision tree and detects the spam emails.The dataset provided to the system is routinely updated so that it detects new type of spams and notifies the user.

5 References

[1] https://www.talosintelligence.com/reputation_center/email_rep

[2] A.K. Sharma and S. Sahni," A Comparative Study of Classification Algorithm for Spam Email Data Analysis", International Journal on Computer Science And Engineering (IJCSE),vol. 3 No. 5 May 2011.

[3] R. K. Kumar, G. Poonkuzhali and P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques", *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol I, IMECS, March 14-16,2012.

[4] A.Chharia and R. K. Gupta, "Email Classifier: an Ensemble using Probability and Rule", IEEE,2013.

[5] A. Iyer, A. Pandey, D. P. K. Pathak and Prof. Mrs. J. Hajgude, "Email Filtering and Analysis using Classification Algorithm*", IJCSI International Journal of Computer Science Issues*, Vol. 11, Issue 4, No 1, July 2014.

[6] H. Kaur and A. Sharma, "Improved Email Spam Classification Method using Integrated Particle Swarm Optimization and Decision Tree", *2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India* 14-16 October 2016.

[7] Enron Dataset, "http://www2.aueb.gr/users/ion/data/enron-spam/