ISSN 2319 - 6629

Volume 8, No.3, April - May 2019

International Journal of Wireless Communications and Networking Technologies

Available Online at http://warse.org/IJWCNT/static/pdf/file/ijwcnt04832019.pdf

https://doi.org/10.30534/ijwcnt/2019/04832019

Email Security Classification of Imbalanced Data Using Naive Bayes Classifier

Alphy Abraham[1], Elice Roy Kanjamala[2], Elsa Mary Thomas[3], Akhila G.P[4]

[1]APJ Abdul Kalam Technological University, India, alphyabraham@cs.ajce.in
[2]APJ Abdul Kalam Technological University, India, eliceroykanjamala@cs.ajce.in
[3]APJ Abdul Kalam Technological University, India, elsamarythomas@cs.ajce.in
[4]Assistant Professor, Amal Jyothi College Of Engineering, India, gpakhila@amaljyothi.ac.in

ABSTRACT

Email is a widely accepted communication method even in large cooperates. However, it is not a safe platform for the companies due to the risk of information leakage, spamming or privacy threat. Corporate-based email systems have different standard etiquette's. It should ensure that the personal information of a company is not being leaked. It is a very hard task for the cyber team to manually check and verify the outbound emails from a company that the private information or important documents of the company is kept safe. So, in the present work we introduce a system to scrutinize the number of outbound emails that the manager has to check by classifying them according to their security levels. This inturn curtails the effort for the cyber security team within the company. The email dataset of all the employees within the company is first obtained in the form of a csv file and is then fed into the classifier to train them and classify new mails according to their security levels. The emails are pre-processed first before classifying them in real time. The high security mails are then send to the manager for verification. And the mails are either forwarded or discarded.

Key words: Classifier, E-mail, Naive Bayes, Security, Text Mining.

1. INTRODUCTION

Electronic mail is a method of exchanging messages between people using electronic devices. The aim of this paper is to show the problems of emails in business sector and solutions for it. The large number of emails has brought some difficulties in large companies, such as spamming, privacy threat or information leakage. As a result, the outbound emails send by the employees must be scrutinized by managers so that the sensitive information will be kept safely without being exposed to the public or other competitive companies. It is not easy to check and verify each and every email by one person or a security team due to several reasons such as time, privacy, efficiency. So, this work suggests a solution to the above mentioned problem with the help of Naive Bayes classifier. It classifies the outbound emails into two categories i.e. high security and low security depending upon the content of the emails.

However, there are practically some issues needed to be solved:

• Due to the privacy policy of the corporation (subjects), we cannot use the metadata of emails. Therefore, the only available information to classify emails is text.

•The portion of sensitive email data in the real world is much less than those of usual email data, which results in data imbalance problem.

•Enormous amounts of training data is required to train the data properly.

In order to solve these major problems, they proposed some methods as follow:

•Extract the meaningful textual features of the email body and the attachments.

•On the issue of imbalanced security labels, use K-means clustering to do under-sampling. The cluster-based majority under-sampling can effectively avoid the critical information loss of majority class.

2. LITERATURE SURVEY

2.1 Security Classification Using ANN

In [1], Jen-Wei Huang explains a method of classifying emails according to three security levels, namely I, II, III using ANN based on the content and the amount of confidentiality in each mail. The body and the attachments of each email forms the data part. Metadata for the email is hidden. The data is stored as an eml file in HDFS. They are loaded and the jobs are partitioned for each executor in the worker node later. The data stored as eml is pre-processed to remove stopwords, such as 'a','and','but', for word segmentation and for stemming. Words are converted to bi-grams which enhances the capability of representing the semantic meaning. The next step in this classification process involves solving data imbalance, training neural network, extract features from the text.ANN is used as classifier to predict the security levels of emails. The power of ANN depends on the amount of available data. Under-sampling is done next to overcome the problem of imbalanced data. An email representation is formed as part of the training process. New emails are fed into the classifier to be classified according to the security levels.

2.2 Cluster-based under sampling for imbalanced Data

An unsupervised learning technique for under sampling through cluster based approaches for majority class problem is described in [2] which will select a representative subset from majority classes and all minority classes. Further takes the samples of training data to improve prediction rates. Class imbalance problem will affect the performance of classifier. Making a correct prediction by the classifier will reduce the cost and to solve the real application problems. Minority class prediction from unbalanced binary class is also mentioned in this work. In this, it explains two strategies K-means algorithm and random sampling approach for majority undersampling and to attain good prediction of minority class. Firstly divide the majority classes into k clusters. To evaluate the accuracy for classification and to know the performance, this work uses Recall, Precision, F-measure, G-mean and BACC (balanced accuracy). Recall will give the accuracy of the predicted positive samples that is the minority samples. Prediction will give the trade- off between accuracy of the predicted negative sample and true positives. In unbalanced data set higher the recall rate and lower will be the precision. Geometric Mean(GM) to maximize the accuracy of positive and negative samples in a favorable manner . There are two approaches for cluster based majority under-sampling prediction (CBMP).

First is learning the classifier by performing random under-sampling on each majority class cluster. Second is by selecting each sample that is closer to the centroid of cluster. Each dataset will perform 5 fold cross validation in which the classification is performed 10 times for the randomness in each cluster. When comparing k-nearest neighbor with the Naive Bayes classifier, k-nearest neighbor is best. From the results, proposed approaches attain high Recall, good Precision, good F-measure, good G-mean, and BACC rates for the imbalanced data.

2.3 KNN approach to unbalanced data distributions

Classification is an important aspect in various applications domains [3]. Many standard algorithms consider that training data are evenly distributed. Standard classification algorithm focuses on the maximization of overall accuracy and thereby ignores the minority class that leads to the overfitting while handling the unbalanced data. The proposed work involves the extraction of protein names where it varies due to multiple naming and short forms from MEDLINE abstract and a learning task on k-nearest neighbors is done. The word level features are extracted for the word representation where each word token is associated with class (protein-start, end, middle) for learning approach. The features include word-token, part-of-speech, and protein-tag.

For the undersampling negative examples in unbalanced data this method uses five different ways which selects the subset of negative examples: random selection, selection of Nearmiss examples (3), and selection of most different examples. NearMiss-1 method selects the negative examples by taking the smallest average distances on three closest positive examples. NearMiss-2 method selection is done by taking the average distances from three farthest positive examples. And NearMiss-3 method selects possible number of closest negative examples for each negative example. In distant negative method, selections of three positive examples that are farthest from the negative are chosen. The random and NearMiss-2 method performs best. The 5NN algorithm is used with the data set and uses Recall, Precision and F-measure for the performance factor. With the increase of negative training examples precision increases, thereby undersampling can be used for adjusting the tradeoff between precision and recall for unbalanced class.

2.4 Security classification using ANFIS

Protection of content depends on the security level of the information and also a challenging one. For this, document should be classified based on the security level using predefined class labels. [4] discuss the classification of documents from TUBITAK UEKAE based on their security level by using ANFIS, SVM [5]s and Naive Bayes. For the classification of textual data pre-processing tasks are involved, where the unstructured documents are converted into structured one. First one is Stemming task for selecting the TF-IDF(term frequency inverse document frequency) values of features to represent the text based document. Second is Feature Selection are used whereby the subset of attributes are selected for modeling the data. It is used to reduce the high dimensionality classification problem that consumes more time. TF-IDF matrix is used for converting unstructured document into structured using feature vector [6], which is holded by the row vector of the document and their weight in each column.

To find the best fitting security level model and to obtain the accuracy two classification algorithms are applied on the training and test data. A cross-validation is done for the parameter selection and classification accuracy. SVM [5] is used for text classification and linearly separable problem and are very standardized and ordinary classifiers. In Naïve Bayes all attributes are independent given the values of the class variable and due to the fuzzy nature of the security level classification this is not offering accurate classification result. ANFIS is used for predicting the security level class labels for text document. The third approach is a hybrid solution onsists of SVM and ANFIS. Documents are classified according to their types and the output of SVM is fed into the input of ANFIS, which classifies the confidential test document. These rates are discretized using the algorithm CACC. This work uses document type as the support vector indicators for detecting the security- level problem.

2.5 Using WordNet to Disambiguate Word Senses for Text Classification

WordNet reflects how human beings organize their lexical memories [7]. The basic building block of WordNet is synset consisting of all the words that express a given concept is an automatic text classification method based on wordsense disambiguation. Disambiguation refers to the removal of ambiguity by making something clear. Disambiguation narrows down the meaning of words and it's a good thing. This word makes sense if you break it down. We use hood algorithm for this disambiguation purpose. Hood is based on the idea that a set of words co-occurring in a document will determine the appropriate senses for one another word despite each individual word being multiply ambiguous. It removes the word ambiguity so that each word is replaced by its sense that is meant in the context. A counter is maintained in each category, which counts the number of words that have its associated senses. The sense of an ambiguous word is determined by the category with the largest counter. Then, the nearest ancestors of the senses of all the non-stop words are selected as the classes of a given document. We apply our "hood" algorithm to Brown Corpus. The effectiveness is evaluated by comparing the classification results with the classification results using manual disambiguation.

Table 1: Comparison of classifiers/techniques

CLASSIFIER	DESCRIPTION	PROS	CONS
Artificial Neural Network	Classifying emails to three security levels, namely I, II,III using ANN.	Ability to learn and model complex relationships.	If the manager is an unauthorised person he might discard the essentials emails also.
Decision Tree	Divides the working area into many subparts.	Flexibility of choosing different subsets of features.	Increases search time and memory space requirements .
KNN	Knn approach to classification problem on extracting information from a large data set.	Undersamplin g is used for adjusting the precision and recall rate for unbalanced data.	2 methods perform best out of five.
Cluster Based Majority Under-samplin g Approach	Selects subset of majority class and all minority class as training data.	Avoid the information loss, improve accuracy attains overall performance.	Increases complexity and time consuming.

3. PROPOSED SYSTEM: EMAIL SECURITY LEVEL CLASSIFICATION SYSTEM

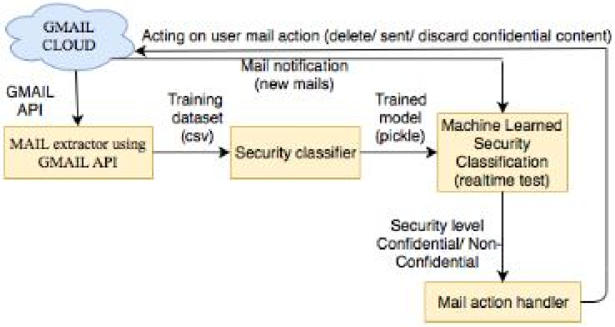
GMAIL CLOUD
Acting on user mail action (delete/ sent/ discard confidential content)
Mail notification (new mails)
GMAIL API
Training dataset (csv)
Trained model (pickle)
MAIL extractor using GMAIL API
Security classifier
Machine Learned Security Classification (realtime test)
Security level Confidential/ Non-Confidential
Mail action handler

Figure 1: Steps for classification

In this section we explain our proposed email classification system. There are several steps to be followed before obtaining a security level for each new email by the classifier. Figure 1 shows the steps for the classification in this work. These include obtaining the email dataset, feeding the data into classifier, processing the cleaned data, fitting the data into training and test datasets etc. and at the end we obtain a security class for the email using Naive Bayes Classifier.

3.1 Naive Bayes Classifier

A classifier is a machine learning model which is used to discriminate different objects based on certain features.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption is that the features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

Types of Naive Bayes Classifier are listed below:
a) Multinomial Naive Bayes: It is used for document classification problem, i.e. whether a document belongs to the category of sports, politics, food, arts etc. The features used by the classifier are the frequency of the words present in the content.
b) Bernoulli Naive Bayes: This is similar to the above method but the features are boolean variables. It predicts only whether it is yes or no. For example to predict whether

apple is a fruit or not.
c) Gaussian Naive Bayes: When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

3.2 Obtain the data set

An Email API (application programming interface) is used to obtain access to emails of different employees in the company, such as generating and sending emails, manipulating templates and enabling access to email metrics. The mails sent and received by different employees are stored in a "CSV" file and categorized as highly confidential and low confidential emails based on the security classification model of the company. In this stage the unstructured data is converted into semi-structural data tables in CSV format. This phase of application is designed as a pluggable module which helps the solution to adapt to each external mail server API's. Primary objective is to extract the valuable information from the mail servers using external API's like GMAIL API, Outlook Mail REST API and lot more. This set of emails is given to the next phase (classifier) for training and testing.

3.3 Training the Classifier

The csv file exported from previous stage is used as input to the classifier. We are going to make use of different packages such as NLTK for processing the messages, matplotlib for visualization, pandas for loading data and NumPy for generating random probabilities for train-test split. To test our model we should split the data into train dataset and test dataset. The train dataset is used to train the model and then it will be tested on the test dataset. We use 75 percent of the dataset as train dataset and the rest as test dataset. Selection of this data is uniformly random. Before starting with training we must pre-process the messages. First, we convert all characters in the message to lowercase. This is because act and ACT mean the same and we do not want to treat them as two separate words. Then we tokenize each message. Tokenization is the task of splitting up a message into pieces, throwing away the punctuation characters.
The words like go, goes, going indicate the same activity. So we replace all these words by go. This process is called stemming. We use a famous stemming algorithm, Porter Stemmer, for this purpose. We then remove the stop words. Stop words are those words which occur frequently. For example words like the, a, an, is, to,' the' etc. These words do not give us any extra information about the context. So we remove those words. We use, yet powerful theorem from probability theory called Bayes Theorem. It is mathematically expressed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Where A and B are events and P (B) is not equal to zero.
• P (A) and P (B) are the probabilities of observing A and B without regard to each other.

• P (A|B) ,conditional probability, is the probability of observing event A given that B true.

• P (B|A) is the probability of observing event B given that A is true.

Using the above theorem we find out the highly confidential data and less confidential data. For classifying the input message, first we pre-process it. For each word w in the processed messaged we find a product of P (w/high). If w does not exist in the train dataset we take TF(w) as 0 and find P (w/high) using above formula. We multiply this product with P (high). The resultant product is the P(high/message). Similarly, we find P (low/message). Whichever probability among these two is greater the corresponding tag (high or low) is assigned to the input message.
Classifier will extract each word in the data table and validate the word with the training model. The classifier fits the data as training and test data and build a classifier model which will predict if a newly arrived email comes under highly confidential category or less confidential category.

3.4 Post Processing

If the classifier classifies an email as highly confidential email, then it has to be send to the managing agent for clarification before sending it to the corresponding recipient or else, it should be ignored from further clarification and be send to the mentioned recipient. If the managing agent (Security team of the company) find that the confidential documents or data of the company is being shared with a competitive opponent company, and then the employee who sends the mail can be accused. If the manager finds that confidentiality of the company is not affected then the manager can forward the email to the corresponding recipient.

4. CONCLUSION

Dataset processed through a classifier would train the email dataset and an label is obtained for a newly arrived email which specifies the category to which the email belongs based on its level of confidentiality. The highly confidential emails are viewed by the security team or the manager of the company and forwarded or discarded to/ from the corresponding recipient.

5. FUTURE WORKS

Currently our project considers emails that doesn't contain subject or body but has attachments as higher confidential emails. In future these attachments can be parsed and checked for confidential contents.

REFERENCES

[1] Jen-Wei Huanga, Chia-Wen Chiang, Jia-Wei Chang (2018). First International Conference on Engineering Applications of Artificial Intelligence 75 (2018) 1121. Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise.https://doi.org/10.1016/j.engappai.2018.07.010

[2] Yan-PingZhang, Li-NaZhang,Yong-Cheng Wang (2010).School Of Computer Science and Technology, Anhui University, Hefei, China Cluster-based Majority Under-Sampling Approaches for Class Imbalance Learning.

[3] Jianping Zhang, Inderjeet Manai,Workshop on Learning from Imbalanced Datasets II, ICML,Washington DC, 2003.KNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction.

[4] Erdem Alparsian,Adem Karahoca, Hayretdin Bahsi. Security-level classification for confidential documents by using adaptive neiro-fuzzy inference systems [2012]

[5] Kunjali Pawar, Madhuri Patil. International Journal Of Computer-Aided Technologies (July 2016).Spam Filtering Security Evaluation Framework Using SVM,LR, MILR.

[6] Mikhail S. Ageev. Support Vector Machine Parameter Optimization for Text Categorization Problems

[7] Y Liu, P Scheuermann, X Li1, X Zhu. Using WordNet to disambiguate word senses for text classification.