I.J. Information Engineering and Electronic Business, 2015, 2, 60-65
Published Online March 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2015.02.08

Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection

Masoumeh Zareapoor
Department of Computer Science, Jamia Hamdard, New Delhi, India
mzarea@jamiahamdard.ac.in

Seeja K. R
Department of Computer Science & Engineering, Indira Gandhi Delhi Technical University for Women, New Delhi, India
krseeja@gmail.com

Abstract—Dimensionality reduction is generally performed when high dimensional data like text are classified. This can be done either by using feature extraction techniques or by using feature selection techniques. This paper analyses which dimension reduction technique is better for classifying text data like emails. Email classification is difficult due to its high dimensional sparse features that affect the generalization performance of classifiers. In phishing email detection, dimensionality reduction techniques are used to keep the most instructive and discriminative features from a collection of emails, consists of both phishing and legitimate, for better detection. Two feature selection techniques - Chi-Square and Information Gain Ratio and two feature extraction techniques – Principal Component Analysis and Latent Semantic Analysis are used for the analysis. It is found that feature extraction techniques offer better performance for the classification, give stable classification results with the different number of features chosen, and robustly keep the performance over time.

Index Terms—Feature Selection, Feature Extraction, Dimensionality Reduction, Text mining, Phishing, Classification.

I. Introduction

Phishing is a new internet crime in comparison with others, such as hacking. The word of phishing is a variation on the word fishing. The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. Phishing is capable of damaging electronic commerce because it causes user to lose their trust on the internet. To make customers aware of latest phishing attacks, some international organizations, such as anti phishing working group (APWG), have published phishing alerts on their websites [1]. According to Anti Phishing working group trends report first quarter 2014 [2], the number of phishing sites increased by 10.7 percent over the fourth quarter of 2013 and also the payment services are the most targeted industry sector. E-mails can be categorized into three [3] - Ham, Spam and Phishing. Ham is solicited and legitimate email while spam is an unsolicited email. On the other hand phishing is an unsolicited, deceitful, and potentially harmful email. Phishing emails are created by fraudulent people to imitate real E-banking emails. Phishing attacks are classified into two [4] as shown in fig.1,

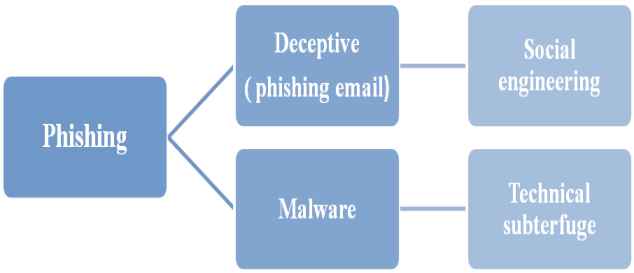

Fig.1. Types of phishing attack

The first one is deceptive phishing which is related to social engineering schemes, depend on forged email that pretence from a legitimate company or bank. Then, through a link within the email, the attacker attempts to mislead users to fake Websites. These fake Web sites are designed to deceptively obtain financial data (usernames, passwords, credit card numbers, and personal information, etc) from genuine users. The second technique is malware phishing that is related to technical subterfuge schemes that rely on malware after users click on a link embedded in the email, or by detecting and using security holes in the user's computer to obtain the his online account information directly. Phishing emails look exactly same as that of e- banking e-mails and they easily traps internet banking users to disclose their banking credentials like bank account number, password, credit card number, and other important information needed for transaction. The attacker then performs fraudulent transaction from the user's account using this collected information.

Several techniques have already been developed for Phishing email detection. They include black listing and white listing [4], network and content based filtering [22], firewalls [4, 21, 22], client side tool bars [4, 21, 22], Server Side filters [21, 22] and user awareness [22]. But the most critical issue with these current techniques is, when classifying email (text), often the data contained in emails are very complex, multidimensional, or represented by a large number of features. This results in high space and time complexity [5] and poor classifier performance. The cost of computing power requirement of the classification algorithms can be reduced by using fewer distinctive features [6]. Thus dimensionality reduction techniques are used for email classification task in order to avoid dimensionality problem. This can be done either by using feature extraction or by using feature selection. In this paper, we used both feature selection and feature extraction techniques, to discriminate between two classes of emails (ham or phishing) by using fewer and more distinctive features, to reduce the computation cost and enhance the results.

The dimensionality reduction techniques like PCA have been popular since the early 90s in text processing tasks [7, 8]. Tsymbal et al. [9] propose two variants of PCA that use the within and between class covariance matrices to take into account the class information. They test the results on typical database data, but not to text categorization. Brutlag and Meek [10] investigate the effect of feature selection by means of common information statistic on email filtering. Xia and Wong [11] discussed the email categorization problem in the context of personal information management.

This paper analyses the effect of various dimensionality reduction techniques in text classification. Feature extraction methods like Principal Component Analysis (PCA) [7] and Latent Semantic Analysis (LSA) [12] are compared with classical feature selection techniques like Chi-Square ($\chi 2$) [13], and Information Gain (IG) [14], which have an established reputation in text classification. In order to study the effectiveness of various dimensionality reduction techniques in phishing email classification, each technique were tested with Bagging classifier [8], which has already proved by researchers, good for e-mail classification.

II. Materials And Methods

A. Dimensionality Reduction Techniques

In text classification tasks, the documents or examples are represented by thousands of tokens, which make the classification problem very hard for many classifiers. Dimensionality reduction is a typical step in text mining, which transform the data representation into a shorter, more compact, and more predictive one [8]. The new space is easier to handle because of its size, and also to carry the most important part of the information needed to distinguish between emails, allowing for the creation of profiles that describe the data set. Two major classes of dimensionality reduction techniques are described in the following sections.

- ***Feature Extraction***

In feature extraction [8], the original feature space is converted to a more compact new space. All the original features are transformed into the new reduced space without deleting them but replacing the original features by a smaller representative set. That is when the number of feature in input data is too large to be processed then the input data will be transformed into a reduced representation set of features.

Principal Components Analysis (PCA)

PCA is a well known technique that can reduce the dimensionality of data by transforming the original attribute space into smaller space. In the other word, the purpose of principle components analysis is to derive new variables that are combinations of the original variables and are uncorrelated. This is achieved by transforming the original variables Y = [y1, y2,..., yp] (where p is number of original variable) to a new set of variables, T = [t1, t2,..., tq] (where q is number of new variables), which are combinations of the original variables. Transformed attributes are framed by first, computing the mean (μ) of the dataset, then covariance matrix of the original attributes is calculated [5]. And the second step is, extracting its eigenvectors. The eigenvectors (principal components) introduce as a linear transformation from the original attribute space to a new space in which attributes are uncorrelated. Eigenvectors can be sorted according to the amount of variation in the original data. The best n eigenvectors (those one with highest eigenvalues) are selected as new features while the rest are discarded.

Latent semantic Analysis (LSA)

LSA method is a novel technique in text classification. Generally, LSA analyzes relationships between a term and concepts contained in an unstructured collection of text. It is called Latent Semantic Analysis, because of its ability to correlate semantically related terms that are latent in a text. LSA produces a set of concepts, which is smaller in size than the original set, related to documents and terms [11, 12]. It uses SVD (Singular Value Decomposing) to identify pat- tern between the terms & concepts contained in the text, and find the relationships between documents. The method commonly referred to as concept searches. It has ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSA is mostly used for page retrieval systems and text clustering purposes. LSA overcomes two of the most problematic keyword queries: multiple words that have similar meanings and words that have more than one meaning.

- ***Feature Selection***

In feature selection technique, a subset of original features is selected, and only the selected features are used for training and testing the classifiers. The removed

features are not used in the computations anymore.

Chi-Square

The Chi-Square ($\chi 2$) [13] is a popular feature selection method that evaluates features individually by computing chi square statistics with respect to the classes. It means that the chi-squared score, analysis the dependency between the term and the class. If the term is independent from the class, then its score is equal to 0, other wise 1. A term with a higher chi-squared score is more informative.

Information Gain

Information Gain [14] is a feature selection technique that can decrease the size of features by computing the value of each attribute and rank the attributes. Then we simply decide a threshold in the metric and keep the attributes with a value over it. It just keeps those top ranking ones. Generally, Information Gain selects the features via scores. This technique can be simpler than the previous one. The basic idea is that we only have to compute the score for each feature that can reflects in discrimination between classes, then the features are sorted according to this score and then just keep those top ranking ones.

B. *Bagging Classifier*

Bagging classifier is an ensemble technique which was proposed by Leo Breiman in 1994. It is designed to improve the stability and accuracy of machine learning algorithms used in classification and regression. The basic principle behind ensemble methods is that a group of "weak learners" can come together to form a "strong learner". Here each individual classifier is a "weak learner", while all the classifiers taken together are a "strong learner". Bagging works by combining classifications of randomly generated training sets to form a final prediction. Such techniques can typically be used as a variance reduction technique by incorporating randomization into its construction procedure and then creating an ensemble out of it. Bagging classifier has attracted much attention, due to its simple implementation and accuracy. Thus, we can call bagging as a "smoothing operation" that has an advantage to improve the predictive performance of regression or classification.

In case of classification, where there are two possible classes {positive, negative}, a classification algorithm creates a classifier on the basis of a training set (in this paper it is email dataset). In the bagging method, it creates a series of classifiers. These classifiers are combined into a "compound classifier". The final prediction of the "compound classifier" is gained from weighted combination of individual classifier predictions. The meaning of this theory can be described as a "voting procedure" where the objective is to find the classifier which is having stronger influence on the final prediction than other classifiers.

III. Phishing E-Mail Classification Frame Work

The phishing email classification frame work used in this research is shown in Fig. 2.

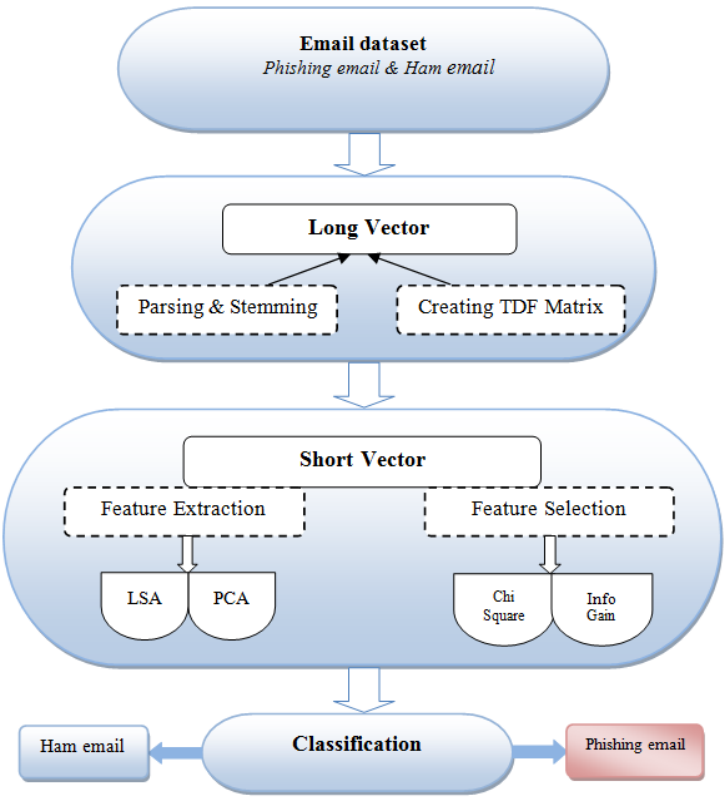

Fig.2. Phishing Email Classification framework

The various steps in phishing email classification are,

- *Data set* is prepared by collecting a group of e-mails from the publicly available corpus of legitimate and phishing e-mails. Then the e-mails are labeled as legitimate and phishing correspondingly.
- *Tokenization* is performed to separate words from the e-mail by using white space (space, tab, newline) as the delimiter.
- Then the words that do not have any significant importance in building the classifier are removed. This is called *stop word removal* and stop words are words like a, the, that etc.
- Then *stemming* is performed to remove in flexional ending from the necessary words.
- Finally, the Term-Document-Frequency (TDF) matrix is created where each row in the matrix corresponds to a document (e-mail) and each column corresponds to a term (word) in the document. Each cell represents the frequency (number of occurrence) of the corresponding word in the corresponding document. Thus, each e-mail in the data set has been converted into an equivalent vector.
- Generally prior to the classification, *dimensionality reduction techniques* are applied to convert the long vector created in step 5. Feature selection or feature extraction techniques are used for dimension reduction and this improves the training time of the classifiers.
- Finally the classification model classifies the dataset into phishing and legitimate.

IV. Implementation

The proposed text mining based text classification is implemented by using the text mining features available in WEKA 3.7.11. The procedure is described in the following sections.

A. Dataset Preparation

Data set is prepared by collecting a group of e-mails from the well known publicly available corpus that most authors in this area have used. Phishing dataset consisting 1,000 phishing emails received from November 2004 to August 2007 provided by Monkey website [15] and, 1,700 Ham email from Spam Assassin project [16]. Then the e-mails are labeled as phishing and legitimate correspondingly.

Table 1. Dataset

Total number of samples	2700
Phishing emails	1000
Legitimate emails	1700

B. Creation of Long Vector

In general, an email consists of two parts: the header and the body message. The header contains information about the message in the form of many fields like sender, subject, receiver, date, etc. The body contains the plain text and may embed within HTML links. In the case of HTML emails, these contain a set of tags to format the text to be displayed on screen. In our work we did not use any separation between body and header. We consider whole the emails itself and so the feature vector contains all the kinds of features like HEADER based feature, URL based feature and BODY based feature [17].

- **Body based feature:** All body-based features occur in the boy of emails and are involved: (body-keyword), (body- jspopup), (body-java script), multipart emails, html emails, verify phrase emails, htmllink, image link and etc.
- **URL based feature:** These features are extracted from the URL link of emails, and included: html-link, url IP ad- dresses, image link and etc.
- **Header based feature:** The features are extracted from the e-mail header like subject, sender, receiver etc.

The extracted features are converted into a long vector by using parsing and stemming. Parsing is a process to extract features from email and analyzing them. Stemming is the process for reducing inflected (or sometimes derived) words to their stem or root form. Then stop words, those words that do not have any significant importance in building the classifiers, are removed. Thus the email dataset of 2700 emails is converted into 2,173 terms (feature). This means that the email dataset can be represented as a term-document matrix with 2700 rows and 2173 columns. Each row in the matrix corresponds to a document (e-mail) and each column corresponds to a term (word) in the document. Each cell represents the frequency (number of occurrence) of the corresponding word in the corresponding document. The vector generated in this stage is considered as long vector.

C. Conversion of Long vector into Short vector

This conversion is done due of 3 reasons:

1. To transform the data representation into a shorter, more compact, and more predictive one.
2. To reduce the complexity of handling features in classification process
3. To increase the speed of classification process

This conversion (Long to Short) can be done either by using Feature Selection or by Feature Extraction. We selected PCA and LSA as feature extraction techniques and Chi-Square and Info Gain as feature selection techniques for the analysis. From the initial 2173 features, small sets of 10,15,50,100,300,500,1000 and 2000 features are selected/extracted with PCA, LSA, Chi-Square and Info Gain, for analysis.

D. Classification

After converting long vector to short vector based on feature extraction and feature selection, we trained

different classifiers on dataset. After several trials and comparison, we decided to use ensemble classifier model bagging with J48 decision tree as the base classifier [18]. The reason for this decision is that, for our dataset and methods (feature extraction and selection), bagging gives good results by reducing the variance of the data set and thus reduces over fitting of the training data.

V. Results And Discussions

The results of various experiments conducted on the selected dataset for different number of features are shown in Fig.3 and Fig.4.

True positive (TP) = number of phishing email that correctly classified as phishing.

False positive (FP) = number of ham email that incorrectly classified as phishing.

For better visualization, the results are presented in the form of the area under the ROC (Receiver Operating Characteristic) curve, which reaches the best value at 1 and worst value at 0. The results are also shown in terms of accuracy of the classification for better understanding.

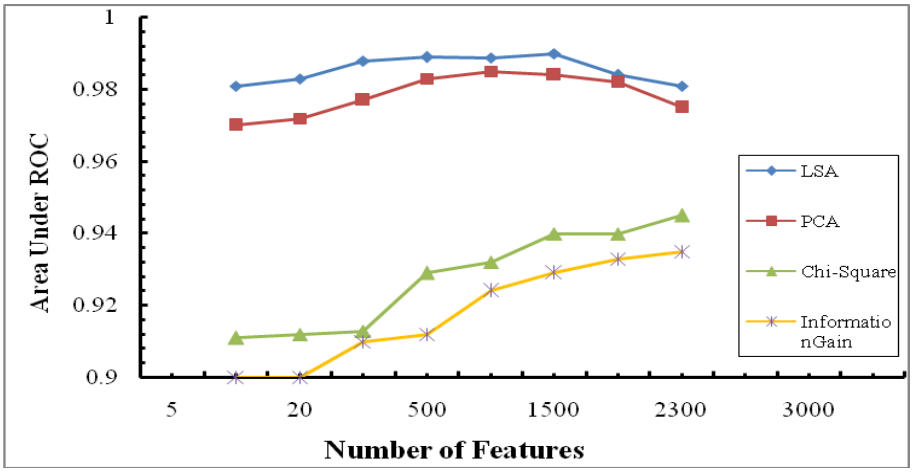

Fig.3. Performance Comparison in terms of ROC Area

It evidents from the Fig (3,4), both FS and FE methods obtain a certifiable results. But FS shows better performance by increasing the number of features, and the results are not better than the FE techniques. On the other hand, the FE methods need commonly much less features to obtain a good performance. In FE methods, choosing more features might degrade the performance of the algorithm. From these results, we can observe that the statistical feature extraction techniques are well suited to discriminate between ham and phishing emails. Especially LSA technique in terms of area under ROC curve shows a good and stable performance irrespective of the num- ber of features chosen.

When we compare the techniques in terms of accuracy (Fig. 4), it is observed that the FE techniques have good performance with a small number of features and the performance values are decreased when a large number of features are chosen, while the FS algorithms need more features for accurate classification. This is because since FS methods directly select features from the dataset, which includes information from the whole dataset, with small number of features they may missed some of the more informative and important features.

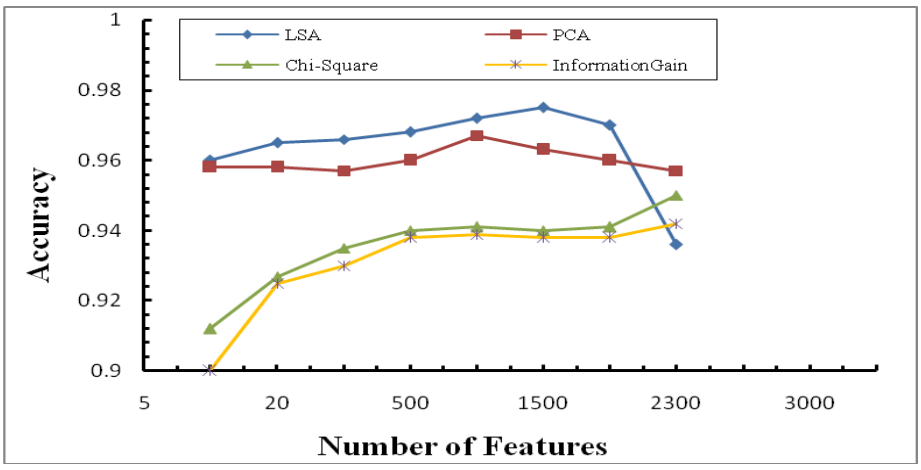

Fig.4. Performance Comparison in terms of Accuracy

VI. Conclusion

In this paper, feature selection methods are compared with statistical feature extraction techniques for email classification. The results show good classification performance when using the feature extraction techniques to classify emails. One of the significant objects in this work is, the results of feature extraction methods (PCA, LSA) are not dependent on number of features chosen. It is an advantage in text classification because choosing the correct number of features in the high dimensional space is a difficult problem. Moreover, Latent Semantic Analysis is found to be the best method, since it outperforms other methods in terms of the area under the ROC curve and accuracy, even when dataset are presented with very few features.

References

[1] APWG. Anti phishing working: http://www.antiphishing.org

[2] Phishing Activity Trends Report 2014: http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf.

[3] I.R.A.Hamid, J.Abawajy. Hybrid feature selection for phishing email detection. International Conference of Algorithms and Architectures for Parallel Processing, (2011), Lecture Notes in Computer Science, Springer, Berlin, Germany; 266-275.

[4] G. L. Huillier, R. Weber, N. Figueroa. Online Phishing Classification Using Adversarial Data Mining and Signaling Games. ACM SIGKDD Explorations Newsletter, (2009), 11(2); 92-99.

[5] J.J. Verbeek. Supervised Feature Extraction for Text Categorization. Tenth Belgian-Dutch Conference on Machine Learning, (2000).

[6] G. Biricik, B. Diri, A.C. Sonmez. Abstract feature extraction for text classification. Turk J Elec Eng & Comp Sci, (2012), 20(1); 1102-1015.

[7] J.C. Gomez, M.F. Moens. PCA document reconstruction for email classification. Computational Statistics and Data Analysis, (2012), 56(3); 741–751.

[8] J.C.Gomez, E. Boiy, M.F.Moens. Highly discriminative statistical features for email classification. Knowledge and Information System, (2012), 31(1); 23-53.

[9] A. Tsymbal, S. Puuronen, M. Pechenizkiy, M. Baumgarten, D.W.Patterson. Eigenvector-based feature extraction for classification. AAAI Press, (2002); 354–358.

[10] J.D.Brutlag, C.Meek. Challenges of the email domain for text classification. In Proceedings of the seventeenth international conference on machine learning, (2000); 103–110.

[11] Y. Xia, K.F. Wong. Binarization approaches to email categorization. In: ICCPOL; 474–481.

[12] G.L.Huillier, A.Hevia, R.Weber, S.Rios. Latent Semantic Analysis and Keyword Extraction for Phishing Classification. IEEE International Conference on Intelligence and Security Informatics, (2010); 129 – 131.

[13] M. Hall, L. Smith. Practical feature subset selection for machine learning. Proceedings of the 21st Australasian Conference on Computer Science. (1998); 181-191.

[14] T. Mori. Information gain ratio as term weight: The case of summarization of IR results. In Proceeding of the 19th international conference on computational linguistics, Taiwan (2002); 688-694.

[15] Phishing Corpus: http://monkey.org/wjose/wiki/doku.php;

[16] SpamAssassin PublicCorpus: http://spamassassin.apache.org/publiccorpus

[17] A. Almomani, T.C.Wan, A.Manasrah, A.Altaher, M.Baklizi, S.Ramadass. An enhanced online phishing e-mail detection framework based on evolving connectionist system. International journal of innovative computing, information and control (2012); **9**(2); 1065-1086.

[18] D. Opitz, R. Maclin. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research**,** (1999), Vol(11); 169-198

[19] F Toolan, J. Carthy. Phishing Detection using Classifier Ensembles. IEEE conference on eCrime Researchers Summit, Tacoma, WA, USA, (2009); 1 – 9.

[20] S.A. Nimeh, D. Nappa, X. Wang, S. Nair. A comparison of machine learning techniques for phishing detection. In Proceedings of the eCrime Researchers Summit, 2007; vol. 1. (Pittsburgh, PA, USA); 60–69.

[21] V. Ramanathan, H. Wechsler. Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation. Journal of Computers & Security, (2013), 34; 123-139.

[22] V.Ramanathan, H.Wechsler. PhishGILLNET-phishing detection methodology using probabilistic latent semantic analysis, AdaBoost and co-training. Journal on information security, 2012.

Authors' Profiles

Masoumeh Zareapoor is a Ph.D. student at Jamia Hamdard University, New Delhi, India. She received her Master degree in computer science from Jamia Hamdard University in 2010.

Seeja.K.R received her Ph.D. degree in Computer Science from Jamia Hamdard University, New Delhi, India, in July 2010. She is currently working as associate professor in the Department of Computer Science & Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India. Her research interests include data mining, algorithm design, bioinformatics and NP-Complete problems.

