iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** X **Month of publication:** October 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Ensemble Model for Classification of Phishing e-mail

Akhilesh Kumar Shrivas[1], Shashibhushan Singh Mahto[2]
Dept. Of IT, Dr. C. V. Raman University

Abstract— Phishing attack is one of the critical issues that access sensitive information from e-mail users like banking password, credit card information and other details. Phishing e-mail not only wastage the storage space in mailbox, decrease the communication band width, but it also damage and misuse the sensitive information. This paper presents the classification of phishing e-mail or non phishing e-mail. In this paper, we have used various classification techniques like C4.5, Classification and Regression Technique (CART), Support Vector Machine (SVM), BayesNet and its ensemble technique for classification of phishing e-mail. The ensemble of CART and SVM gives better actuary results as 99.03% in case of 80-20% training-testing partitions.

Keywords— Phishing, Classification, Ensemble, e-mail, Decision Tree.

I. INTRODUCTION

Now days, security of information is very crucial issue for every organization. Phishing e-mail is one of the important crucial issues for every organization that face by every e-mail users. Data mining based classification techniques play very important role for classification of phishing and non phishing (ham) e-mail. There are various authors have worked in the field of classification of phishing e-mail data. P. Likarish et al. (2008) [1] have suggested B-APT anti phishing tool that is Bayesian Anti-Phishing Toolbar and compared B-APT with Internet Explorer and FireFox. The proposed B-APT tools given better performance than others. J. Yearwood et al. (2010)[3] have used boosting algorithm (AdaBoost) as well as SVM to generate multi-label class predictions on three different datasets created from hyperlink information in phishing e-mails. V. Shreeram et al. (2010) [2] have suggested genetic algorithm for detection of phishing web pages by using rule-based system. I. Rahmi A. Hamid et al. (2011) [4] have analyzed the various models like Bayesian Net, AdaBoost, Decision Tree and Random Forest using phishing data set. Random Forest given highest accuracy as 93% of accuracy. A. Almomani et al. [5] (2012) have discussed various phishing techniques to classify the phishing and non phishing data and they also compared and discussed advantages and disadvantages of various machine learning techniques for phishing e-mail detection and prediction. Andronicus A. Akinyelu et al. (2014) [6] have suggested Random Forest decision tree algorithm for phishing e-mail classification. H. S. Hota et al. (2016) [7] have suggested BBFST-C4.5 model for classification of phishing e-mail which given 98.88% of accuracy with all features.

II. ARCHITECTURE OF PROPOSED SYSTEM

The architecture of proposed system shown in fig. 1. The phishing e-mail data set different partitions into three categories like 60-40%, 75-25% and 80-20% as training-testing. The data set is applied on various techniques C4.5, CART, SVM and bayes net. We have also developed the ensemble models using these individual models and proposed the ensemble of CART and SVM as best classifier for classification of phishing e-mails data. Finally calculate the various performance measures like accuracy, sensitivity and specificity.

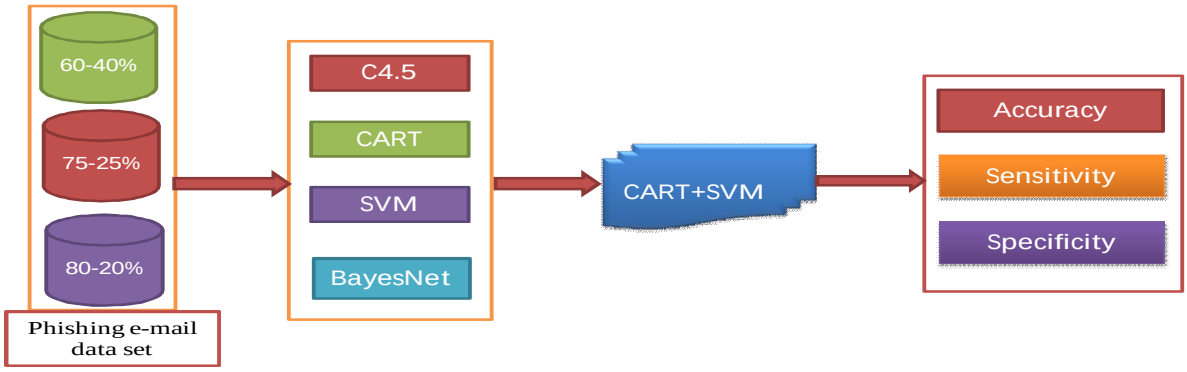


Fig.1: Proposed architecture of phishing e-mail classification

III. METHODS AND MATERIALS

We have used various classification techniques, phishing e-mail data set and WEKA software tools used in this research work as described below:

A. Decision Tree

Decision tree is very popular data mining technique. Decision tree can be handle high dimensional of data .It can be easily converted into classification rule. In this research work, we have used C4.5 and CART technique for classification of phishing e-mail.

C4.5 (Pujari, A. K., 2001) [8] is an extension of ID3 that handle the unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision tree, we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute values are available. We can classify the records that have unknown attribute value by estimating the probability of the various possible results. C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

CART (Classification and Regression Tree) (Pujari, A. K., 2001) [8] is one of the popular data mining techniques of building decision tree. It builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

B. Support Vector Machine (SVM)

Support vector machines (SVMs) (Olson, D. L. et al., 2008) [10] are supervised learning methods that generate input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. SVMs belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. They are also said to belong to “kernel methods”. In addition to its solid mathematical foundation in statistical learning theory, SVMs have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, bioinformatics, face recognition, image processing and text mining, which has established SVMs as one of the most popular, state-of-the-art tools for knowledge discovery and data mining.

C. Bayesian Net

Bayesian classifiers (Han, J. et al., 2006) [9] are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes’ theorem. Classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

D. Phishing e-mail Data Set

This research work used phishing e-mail data set collected from: http://khonji.org website [11]. This data set consist 8266 instances, 48 features and 1 class having phishing and ham. The data set consists 4116 instances and 4150 instances of phishing and non-phishing (ham) respectively.

E. Performance Measures

Various performance measures can be evaluated using some well known statical measures like accuracy, sensitivity and specificity. These measures are calculated by true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Confusion matrix (Han, J., 2006) [9] for two classes are defined in TABLE I. The confusion matrix can be defined by TP, TN,FP and FN. TABLE II shows that equations of various performance measures, where N represents that the total number of samples.

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

TABLE I: CONFUSION MATRIX FOR POSITIVE AND NEGATIVE SAMPLES

Actual Vs. Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

TABLE II: PERFORMANCE MEASURES

Measures	Equation
Accuracy	(TP+TN)/N
Sensitivity	TP/ (TP +FN)
Specificity	TN/ (TN +FP)

IV. RESULT AND DISCUSSION

The experiment carried out using open source WEKA data mining software tool [12] in window environment. In this experiment, we have used various data mining techniques like C4.5, CART, SVM and Bayesian Net for classification of phishing e-mail. We have partitions the phishing e-mail data set into three different training-testing partitions like 60-40%, 75-25% and 80-20% and applied this data set into various models. We have also ensemble two or more models with various combination like C4.5+CART, C4.5+CART+SVAM, CART+SVM , SVM+Bayes Net etc., but we have achieved satisfactory accuracy in case of ensemble of CART and SVM (CART+SVM) as 99.00%,98.98% and 99.03% with 60-40%, 75-25% and 80-20% training-testing data partition respectively. TABLE III shows that accuracy of individuals and ensemble model with different data partitions. Fig. 2 shows that accuracy of model with different data partitions. TABLE IV shows that confusion matrix of best ensemble (proposed) model with different data partitions. In case of 60-40% data partition, 9 and 24 samples are misclassified of ham and phishing e-mail respectively, similarly samples of ham and phishing e-mails are misclassified for other data partitions. TABLE V shows that various performance measures of best model like sensitivity, sensitivity and specificity with different data partitions. Fig. 3 show that performance measures of best model with different data partitions.

TABLE III: ACCURACY OF MODEL WITH DIFFERENT PARTITIONS

Model	60-40%	75-25%	80-20%
C4.5	98.70	98.69	98.67
CART	98.97	98.83	98.97
SVM	98.79	98.83	98.79
BayesNet	97.73	97.91	97.82
CART+SVM	**99.00**	**98.98**	**99.03**

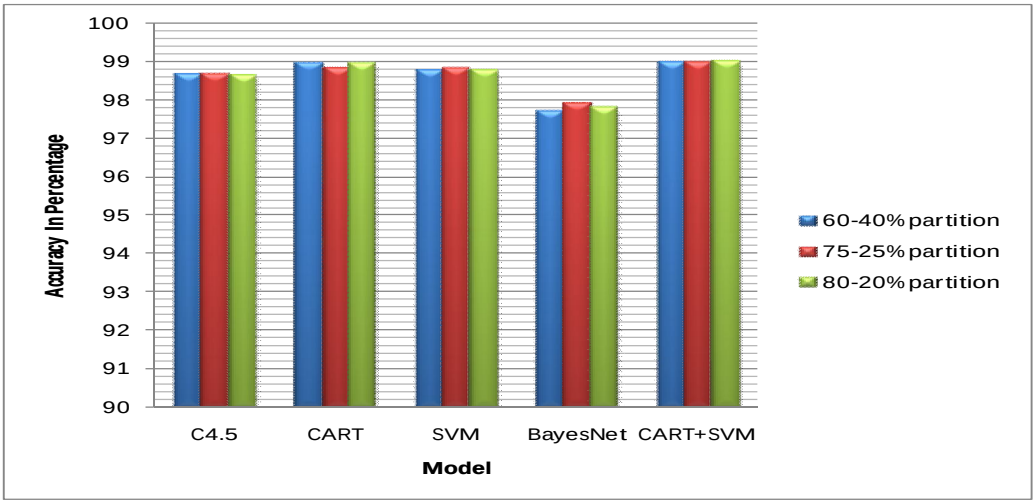


Fig. 2 Accuracy of models with different data partitions

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

TABBLE IV: CONFUSION MATRIX OF BEST MODEL (CART+SVM)

Actual Vs. Predicted	60-40% partition		75-25% partition		80-20% partition	
	Ham	Phishing	Ham	Phishing	Ham	Phishing
Ham	1641	9	1018	9	833	6
Phishing	24	1632	12	1027	10	804

TABLE V: PERFORMANCE MEASURES OF MODEL (CART+SVM)

Models	60-40% partition	75-25% partition	80-20% partition
Accuracy	99.00	98.98	99.03
Sensitivity	99.45	99.12	99.28
Specificity	98.55	98.84	98.77

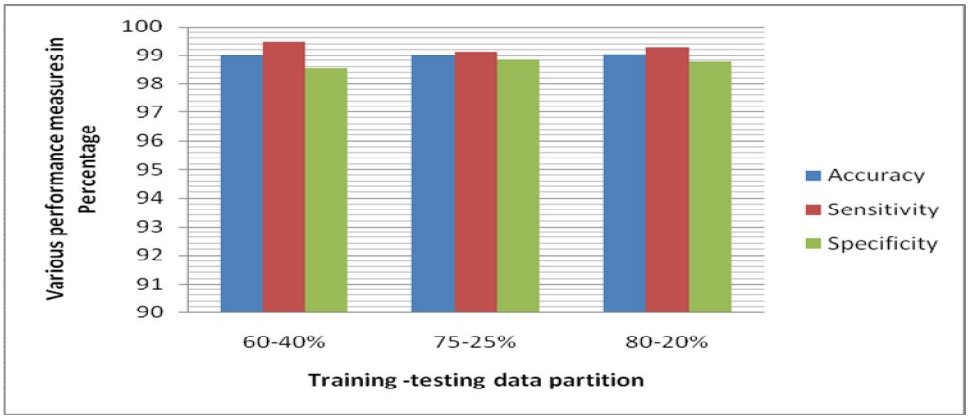


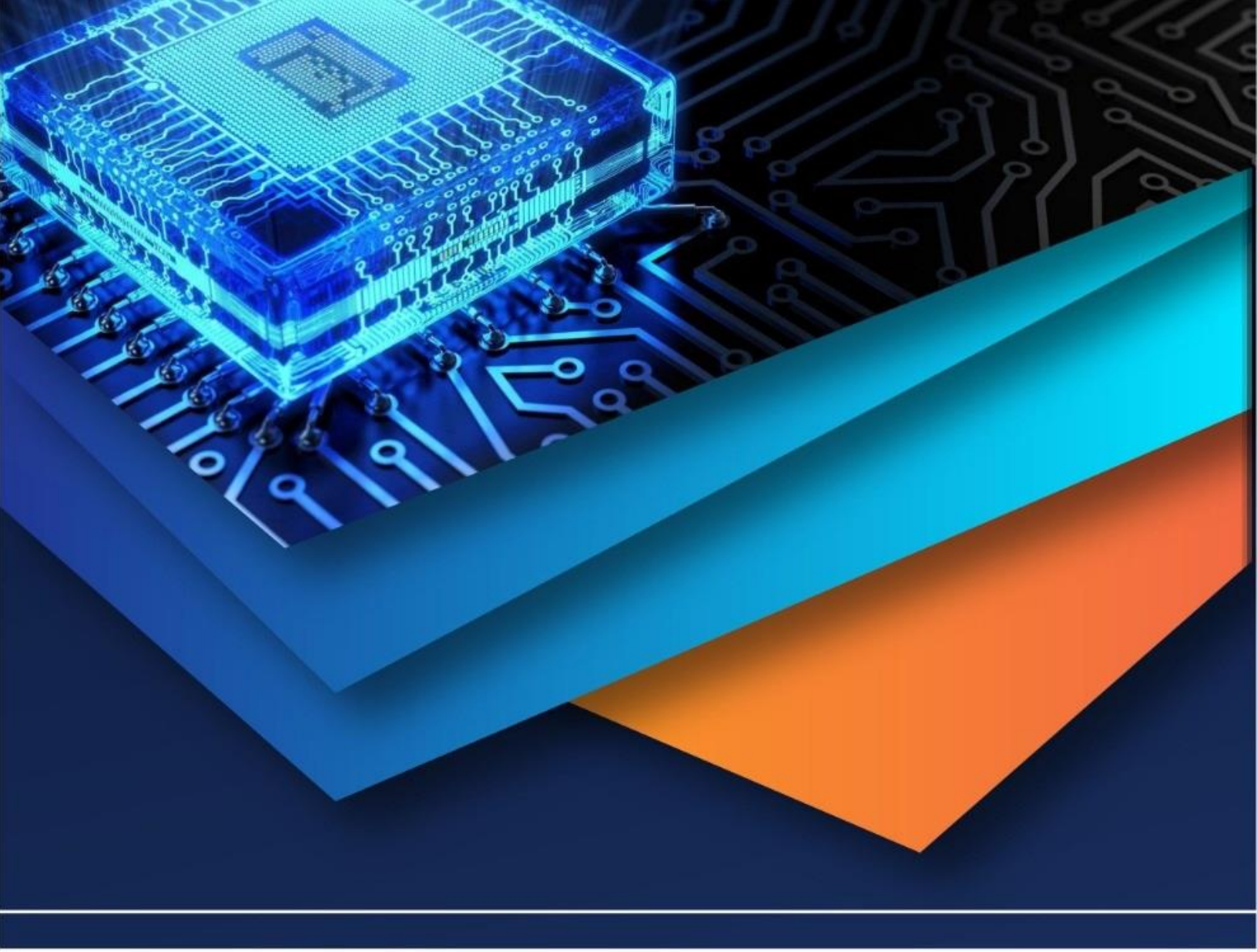
Fig. 3: Various performance measures

V. CONCLUSIONS

Phishing attacks are very serious threat that is faced by every e-mail users. The solution of phishing attack problem is that, the user should not blindly believe any website which enters the sensitive information from users like password. The proposed approach of this paper is to develop the robust model for classification of phishing e-mail. The partitions of data play important role in classification accuracy, because accuracy is changing from partition to partition. The proposed ensemble of CART and SVM gives better accuracy as 99.03% of accuracy in case of 80-20% training-testing data partition for classification of phishing e-mail. In future, we will apply feature selection technique to computationally increase the performance of model.

REFERENCES

[1] P. Likarish, D. Dunbar and T. E. Hansen, "B-APT: Bayesian Anti-Phishing Toolbar", IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings, 2008.
[2] V. Shreeram, M. Suban, P. Shanthi and K. Manjula, "Anti-phishing Detection of Phishing Attacks using Genetic Algorithm", IEEE, 2010.
[3] J. Yearwood, M. Mammadov and A. Banerjee, "Profiling Phishing Emails Based on Hyperlink Information", 2010 International Conference on Advances in Social Networks Analysis and Mining, DOI: 10.1109/ASONAM.2010.56, 2010.
[4] I. Rahmi A Hamid and J. Abawajy, "Phishing Email Feature Selection Approach", 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11, DOI: 10.1109/TrustCom.2011.126, 2011.
[5] A. Almomani, T. C. Wan, A. Manasrah, A. Altaher, E. Almomani, K. Al-Saedi, A. ALnajjar and S. Ramadass, "A survey of Learning Based Techniques of Phishing Email Filtering", International Journal of Digital Content Technology and its Applications(JDCTA) ,Vol. 6,No. 18, 2012.
[6] A. A. Akinyelu and A. O. Adewumi, "Classification of Phishing E mail Using Random Forest Machine Learning Technique", Journal of Applied Mathematics, Vol. 2014, pp. 1-6, 2014.
[7] H. S. Hota, A. K. Shrivas and R. Hota, "A Proposed Backet Based Feature Selection Technique (BBFST) for classification of phishing E-mail classification",4th International Conference on Advance Computing, Networking and Informatics (ICACNI 2016)", NIT Rourkela, Orisha, held on 22-24 , 2016.
[8] A. K. Pujari, "Data Mining Techniques, Universities Press (India) Private Limited", 4th ed., ISBN: 81-7371-380-4, 2001.
[9] J. Han, and M. Kamber, "Data Mining Concepts and Techniques. Morgan Kaufmann, San Francisco", 2nd ed., ISBN: 13: 978-1-55860-901-3, 2006.
[10] Olson, D. L. and Delen, D., Advanced Data Mining Techniques. USA, Springer Publishing: ISBN: 978-3-540-76916-3, 2008.
[11] Web source: http://khonji.org/phishing_studies.html last accessed on July, 2016.
[12] Web source: http:// www.cs.waikato.ac.nz/~ml/weka/ last accessed on June, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY