UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 1 | Year: 2016 | ISBN: 978-967-0194-81-3

A Comparative Study Between Support Vector Machine And Neural Network For Phishing E-Mail Detection

Syahir Mohammad[1], Noorfa Haszlina Mustaffa*[2]

Department of Computer Science, Faculty of Computing,
Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia
[1]syahirmohammad86@gmail.com, [2]noorfa@utm.my

Abstract

In globalization era, email is a very useful connectivity to users around the world. Apart from email to communicate between them, an email is also easier to use, faster and cheaper. The problem of email is the existence of e-mail phishing. The phishing e-mail is unknown from where the email was sent and it usually exists in advertising. Many researchers have studied to solve the problem of phishing emails. The techniques have been proposed is machine learning to detect the phishing e-mail by used classification of dataset. Therefore, in this study, the main objective is to use the classifier method it is Support Vector Machine and Neural Network in order to process each of dataset for classify the data it is phishing or non-phishing e-mail because to evaluate the accuracy of classification. Once result train have process by kernel Radial Basis Function (RBF) then produced best value of parameter it will used test data for produced accuracy result using Support Vector Machines (SVM). Initially, phishing e-mail will be test using the full set of data it is full features. It will be test by SVM. Those results will be compared with other classifier it is Support Vector Machine and Neural Network.. The result show Support Vector Machine is produced highest accuracy of classification than Neural Network it means this technique can be help for phishing e-mail detection. The Neural Network also have produced better accuracy result, it means this technique can used for classification for future work. Last but not least, the result also measured to an accuracy of classification.

Keywords: Phishing, Support Vector Machine, e-mail phishing, classification, Neural Network

1.0 Introduction

All activity of widespread usage of Internet for online banking and trade, phishing attack and allied form of identity theft that will more becoming extremely and more dangerous popular among the hacker communication. The anonymity in the Internet, coupled with the potential for large financial gains, serves as strong motivation to perpetrate such seeming low risk, yet high return crimes. This is initial way for the wrong hand to gain the information from others to something that really illegal that make the other person will lose everything such as of account banking money and confidential data

Most of the recent phishing attacts are proceed with certain step to success their process of phishing. In the first step, the phishers harvest the e-mail addresses of their possible victims from social engineering attack such as webpages and forum. Then, large volume of phishing e-mails impersonating legal banking domains are sent out using anonymous SMTP servers or compromised machine. The e-mail contain hyperlink to lure the recipients into a camouflage website with the appearance similar to the real website to gain personal information from victim that have weakness of identify of camouflage of website.

The fake website contains input forms requesting personal critical information such as credit card, social security number, mother's maiden name, contact number etc. Although exiting of spam filtering techniques can employed to combat phishing emails, there are also skill for countermeasures to prevent phishing, then this is also not entirely scalable as there are a vast number of readily available tools that can bypass both the statistical and rule based spam filters. As these mechanisms are not completely tuned for detection of phishing e-mails are prevalent. Furthermore, unlike spamming which impacts bandwidth, phishing attack directly affect their victims by inflicting heavy loss due to monetary damage.

Several browser extensions and plug-ins have been proposed to address the problem of phishing attacks. Although these techniques are partially effective in determining the authenticity of visited website, they suffer from one or more limitation. First, as these approaches operate on the camouflage website they will expose users to one step closer to the attacker website. Secondly, as these tools detect phishing attacks based on the legitimacy of the domain address (IP), they fail to protect the users when the attacker is launched by used legitimate domain. For example, an attacker could compromise the web server and then launch the fake pages/ pop-up within the context of legitimate domain. Also, one of the recent phishing attacks targeted on Yahoo via its web hosting domain geocities.com had the attacker create a username 'login' and a login launch page similar to the geocities authentication page appearing as www.geocitis.com/login (note that here 'login' refers to the username).

The objectives of this study are : (1) to propose comparative study a phishing e-mail detection use Support Vector Machine and Neural Network method in order to compare and evaluate result accuracy between this machine learning, (2) to implement and apply Support Vector Machine technique for produced result accuracy classification phishing e-mail, (3) to implement and apply Neural Network technique for classify phishing e-mail and produced result accuracy classification, and (4) to evaluate and compare the performance model SVM and NN in accuracy of detection phishing e-mail.

2.0 Problem Statement

There are many of phishing e-mails have distribute in inbox user and make that content interested to the victim because attacker will make that content related to victim, this really difficult for victim to identify the fake e-mails and attacker will provide fake website to fill in the personal information because most of user will lack of to classify the phishing e-mail. There are many techniques used in order to detect phishing e-mail. Some of them have the limitation which includes the less of accuracy. In this comparative study, will be used different classifier it is SVM and NN to evaluate the accuracy and efficiency of phishing e-mails detection.

3.0 Methodology

Methodology is really important thing need to focus for to implement of project to more systematic and it will follow by time frame of plan that have been discuss it before. It is set of important technique when want to solve problem, it is really needed if have any of problem some of internal or external. The methodology will use in this paper it is combine of classification (Support Vector Machine) and Bee Algorithm, both of this approach it will use for detect what kind of phishing e-mail will come before reach in inbox user. This is for review of process flow.

The RBF kernel function is really important for accuracy of classification result that may run continuously for produced the best result. It will have the vector of input into the classification as linear that can be achieve. The kernel function is used for as usual linear kernel then SVM have parameter C that could be used for the point have problem to classify as linear. This problem will happen then it has solution to used kernel method in produced best value of parameter in. The dataset used division 70% and 30% with full attributes for make the comparable of study in this experiment.

i. Executed in RBF Kernel

The function kernel used in SVM algorithm, for this kernel will be used is Radial Basis Function kernel that have in LIBSVM toolbox. The function of this kernel used two parameter that be considered, for value parameter is cost C and parameter gamma. The function of this kernel will use for train data then produced the result, from that phase it will be used for test data for produced accuracy of classification.

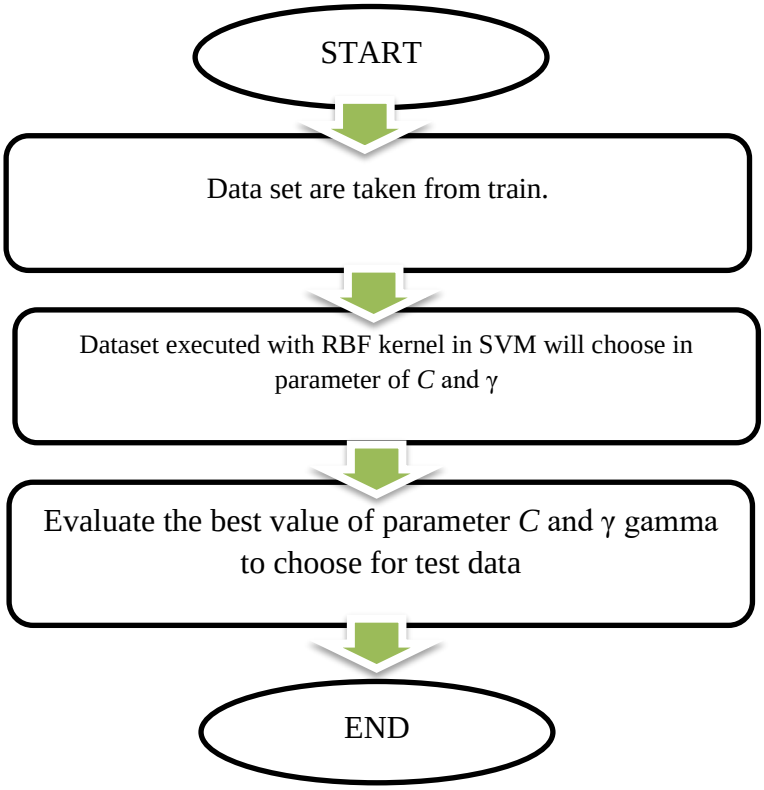

Figure 1 Flowchart of phase executed in RBF kernel process

ii. Executed in Neural Network

The process of NN will executed with full features in dataset then has division of data it is 70% and 30% for make the comparison of both technique with comparable study in result classification. From train data, it will process data for evaluate the accuracy that have in data. Therefore it is important to process test data in order for evaluate the result classification, in test data process it should important in setup parameter value hidden layer, learning time and momentum.

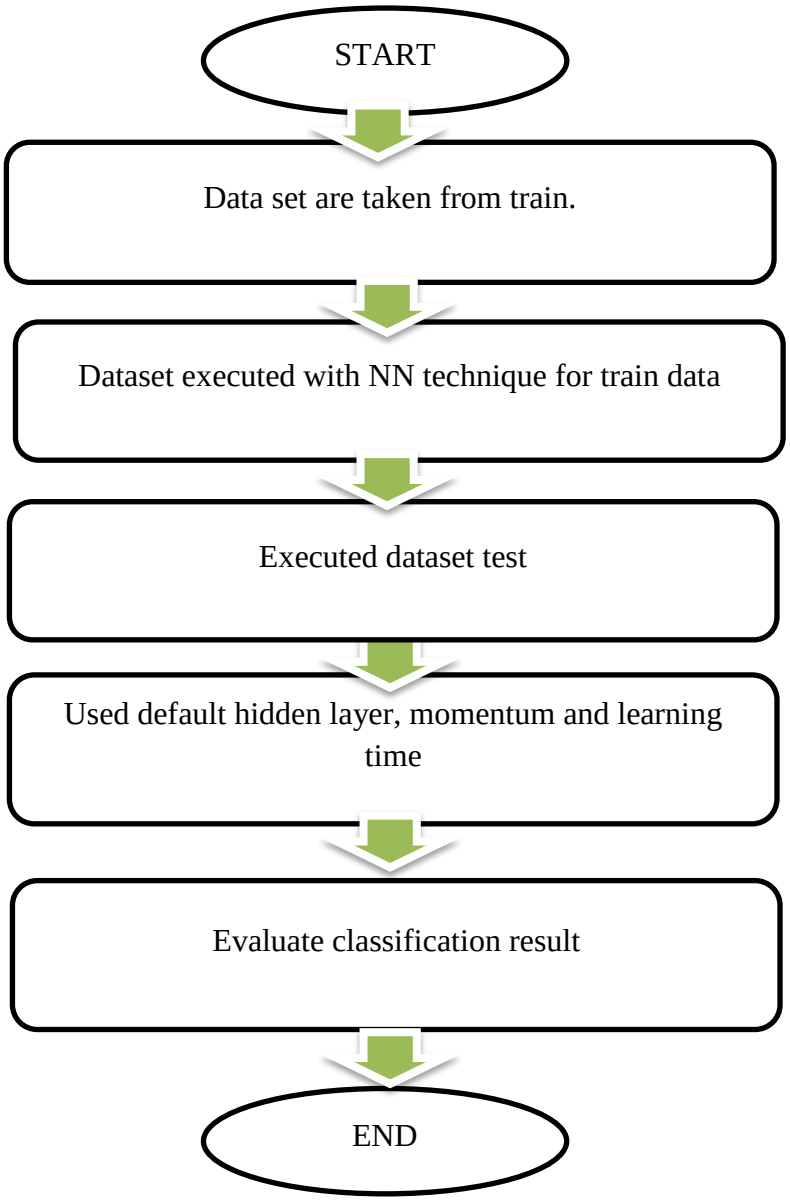

Figure 2 Flowchart of NN

For the configuration it is set as default then produced the result of accuracy for analysis the technique in classification. The parameter of value it is can be choose the best value for produced result accuracy classification in order for make the highest accuracy percentage.

iii. Evaluate and Compare Result Accuracy SVM and NN

For the last phase it is for evaluation both technique it to compare classifier model or method which is about the accuracy percentage result. The result produced by NN and SVM used different value of parameter to produced best accuracy result.

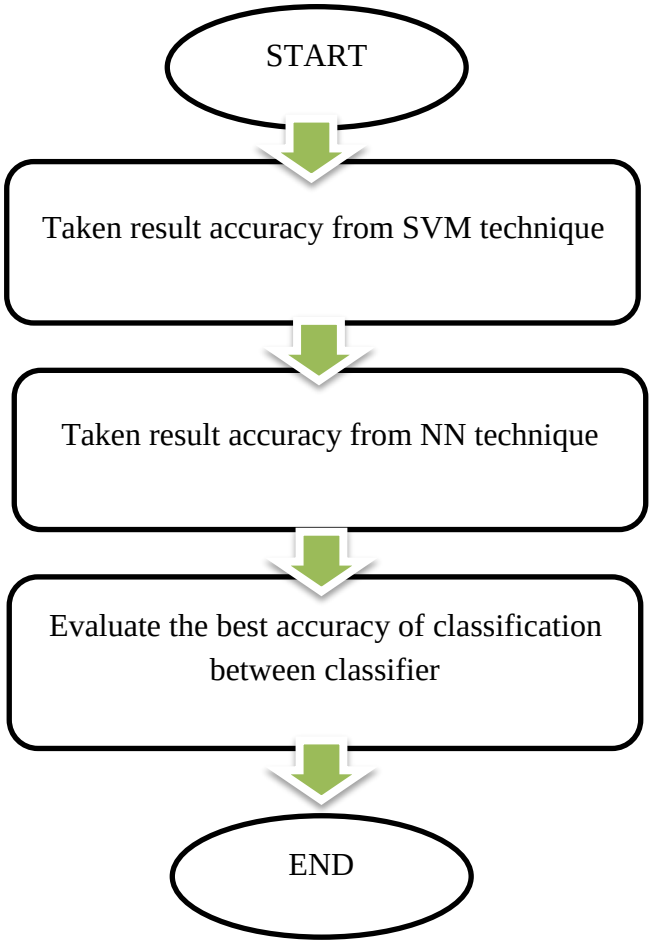

Figure 3 Flow evaluation result between NN and SVM.

3.1 Dataset Information

Based on the link that given that provide by are Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt. The information of dataset is taken from UCI Machine Learning Repository. Table 1 is about information the dataset. This data were classified the email as phishing e-mails (1) and non-phishing (0) at the end of the final attribute. From above that content in Table 3.1, the attribute involve is 57. The run-length attributes measure the length of sequences of consecutive capital letters..

Table 1 : Information Dataset

Dataset characteristic:	Multivariate	Number of attribute:	57
Attribute characteristic	Integer, real	Associated task:	Classification
Number of instances	4601	Area:	Computer

4.0 Result

The result show in Table 2 after the classification process is done with full features based on the accuracy produced by SVM and NN classifier. The training data use only 3221 with have all phishing and non-phishing emails. The test data use 1380 dataset with have full features that combined phishing and non-phishing data. This section will show result of test data then the best value of *C* and *y* will selected after the accuracy of train data have show result. The result of train and test are show in Table 2 and Table 3.

Table 2 Result accuracy train data

Dataset Full Features	**Accuracy**
Train Data 70% SVM	93.16%
Train Data 70% NN	90.36%

Table 3 Result accuracy test data

Dataset Full Features	**Accuracy**
Test Data 30% SVM	86.73%
Test Data 30% NN	86.08%

From the experiment use SVM method with kernel RBF it show have highest accuracy for division dataset 70% and 30% followed by 60% and 40% have lowest accuracy. The division dataset 70% and 30% have low accuracy compare to 90% and 10%. Therefore, division dataset 90% and 10% have highest accuracy compare to 60% and 40%.

Table 4: Result Accuracy SVM with Different Division data

Dataset Full Features	Type Division dataset	Division dataset (%)	Default C= 1 γ =1 Result Accuracy (%)	Best Value of Paramater	Accuracy (%)
57	Train	70	93.16	C= 1024 γ =0.00012207	86.73
	Test	30			
	Train	60	78.20	C = 128 γ = 0.25	78.15
	Test	40			
	Train	90	92.39	C = 512 γ = 0.000244141	87.4
	Test	10			

The Table 5 show result with different division of dataset for this experiment in order to make conclusion what will happen when use the different division of data used NN technique.

Table 5: Result Accuracy NN with Different Division data

Dataset Full Features	Type Division dataset	Division dataset (%)	Accuracy (%)
57	Train	70	86.08
	Test	30	
	Train	60	80.19
	Test	40	
	Train	90	84.78
	Test	10	

In Table 6 show time taken for different division of dataset use SVM technique, it is to show how long time taken to process classification of data in order to produced accuracy result percentage.

Table 6 : Time taken for accuracy data in SVM

Division of Dataset	Train Time (s) (SVM)	Train Time (s) (NN)	Test Time (s) (SVM)	Test Time (s) (NN)	Accuracy (%) (SVM)	Accuracy (%) (NN)
70% and 30%	15m.30s	14m.61s	0m.8s	0m.10s	86.73	86.08
60% and 40%	15m.01s	14m.30s	0m.9s	0m.13s	78.15	80.19
90% and 10%	20m.20s	17m.50s	0m.7s	0m.10s	87.4	84.78

5.0 Discussion

From this section will discuss about comparison with other classifier which is use in SVM and NN technique. Both technique in this experiment used the same dataset is from UC Irvine Repository. The main objective in this experiment is apply both method it is Support Vector Machine and Neural Network for classify the phishing e-mail. From that method have produced result of accuracy classification then compared both classifiers to evaluate the accuracy of classification.

6.0 Conclusion

In conclusion, objective has been mentioned before in Chapter 1, this project had fulfilled all three objectives or the requirement that had stated. From all this full experiment, by using the method of Support Vector Machine is a good algorithm to improve find out result of accuracy classification performance. Hopefully, this project will contribute with anyone or student to further more study to improve about classification on this field for detect phishing and non-phishing e-mail.

References

Almomani, A., Wan, T. C., Altaher, A., Manasrah, A., ALmomani, E., Anbar & Ramadass, S. (2012). Evolving fuzzy neural network for phishing emails detection. Journal of Computer Science, 8(7), 1099.

Zhang, N., & Yuan, Y. (2013). Phishing detection using neural network.Department of Computer Science, Department of Statistics, Stanford University. Web, 29.

Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012, June). The livehoods project: Utilizing social media to understand the dynamics of a city. In International AAAI Conference on Weblogs and Social Media (p. 58).

Zhang, C., Chen, W. B., Chen, X., & Warner, G. (2009, March). Revealing common sources of image spam by unsupervised clustering with visual features. In Proceedings of the 2009 ACM symposium on Applied Computing(pp. 891 892). ACM.

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007, October). A comparison of machine learning techniques for phishing detection. InProceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (pp. 60-69). ACM.

Chandrasekaran, M., Narayanan, K., & Upadhyaya, S. (2006, June). Phishing emaildetection based on structural properties. In NYS Cyber Security Conference (pp. 1-7).

Toolan, F., & Carthy, J. (2009, September). Phishing detection using classifier ensembles. In eCrime Researchers Summit, 2009. eCRIME'09.(pp. 1-9). IEEE.

Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010).New filtering approaches for phishing email. Journal of computer security, 18(1), 7-35.

Aburrous, M., Hossain, M. A., Thabatah, F., & Dahal, K. (2008, April). Intelligent phishing website detection system using fuzzy techniques. InInformation and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on (pp. 1-6). IEEE.

Chu, W., Zhu, B. B., Xue, F., Guan, X., & Cai, Z. (2013, June). Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs. In Communications (ICC), 2013 IEEE International Conference on (pp. 1990 1994). IEEE.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010, April). Predicting phishing websites using classification mining techniques with experimental case studies. In Information Technology: New Generations (ITNG), 2010 Seventh International Conference on (pp. 176-181). IEEE.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. Expert systems with applications, 37(12), 7913-7921.

Hamid, I. R. A., & Abawajy, J. (2011). Hybrid feature selection for phishing email detection. In Algorithms and Architectures for Parallel Processing (pp. 266 275). Springer Berlin Heidelberg.