{tag}

{/tag}
International Journal of Computer Applications

Volume 122 - Number 20

Year of Publication: 2015

Authors:
Doaa Hassan

10.5120/21813-5191

{bibtex}pxc3905191.bib{/bibtex}

Abstract

Phishing websites are a form of mimicking the legitimate ones for the purpose of stealing user 's confidential information such as usernames, passwords and credit card information. Recently machine learning and data mining techniques have been a promising approach for detection of phishing websites by distinguishing between phishing and legitimate ones. The detection process in this approach is preceded by extracting various features from a website dataset to train the classifier to correctly identify phishing sites. However, not all extracted features are effective in classification or equivalent in their contribution to its performance. In this paper, we investigate the effect of feature selection on the performance of classification for predicting phishing sites. We evaluate various machine learning algorithms using a number of feature subsets selected from an extracted feature set by various feature selection techniques in order to determine the most effective subset of features that results in best classification performance. Empirical results shows that using our new proposed methodology for selecting features by removing redundant ones that equally contribute to the classification accuracy, the decision tree classifier achieves the best performance with an overall accuracy of 95. 40%, false positive rate (FPR) of 0. 046 and false negative rate (FNR) of 0. 065.

References

- http://www. support-vector-machines. org/.
- https://archive. ics. uci. edu/ml/datasets/Phishing+Websites.
- Ram B. Basnet, Andrew H. Sung, and Quingzhong Liu. Feature selection for improved phishing detection. In Proceedings of the 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems: Advanced Research in Applied Artificial Intelligence, IEA/AIE'12, pages 252–261, 2012.
- Andr Bergholz, Gerhard Paa, Frank Reichartz, Siehyun Strobel, and Schlo Birlinghoven. Improved phishing detection using model-based features. In Fifth Conference on Email and Anti-Spam, CEAS, 2008.
- Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. Lexical feature based phishing url detection using online learning. In Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, AISec '10, pages 54–60, 2010.
- Weibo Chu, Bin B. Zhu, Feng Xue, Xiaohong Guan, and Zhongmin Cai. Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing urls. In Proceedings of IEEE International Conference on Communications, ICC 2013, Budapest, Hungary, June 9-13, 2013, pages 1990–1994, 2013.
- M. Dash and H. Liu. Feature selection for classification. Intelligent Data Analysis, 1:131–156, 1997.
- Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In Proceedings of the 16th International Conference on World Wide Web, WWW '07, pages 649–656, 2007.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. SIGKDD Explor. Newsl. , 11(1):10–18, November 2009.
- A. G. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker. On the Relationship Between Feature Selection and Classification Accuracy. In JMLR: Workshop and Conference Proceedings 4, pages 90–105, 2008.
- Esra Mahsereci Karabulut, Selma Aye zel, and Turgay briki. A comparative study on the effect of feature selection on classification accuracy. Procedia Technology, 1(0):323 – 327, 2012. First World Conference on Innovation and Computer Sciences (INSODE 2011).
- Ron Kohavi and George H. John. Wrappers for feature subset selection. Artif. Intell. , 97(1-2):273–324, December 1997.
- L. Ladha and T. Deepa. Features selection methods and algorithms. International Journal on Computer Science and Engineering (IJCSE), 3(5):1787–1797, May 2011.
- Chang-Hwan Lee, Fernando Gutierrez, and Dejing Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining, ICDM '11, pages 1146–1151, 2011.
- K. Ming Leung. Naive bayesian classifier, 2007. Department of Computer Science / Finance and Risk Engineering- Polytechnic University.
- Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Beyond blacklists: Learning to detect malicious web sites from suspicious urls. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pages 1245–1254, 2009.
- Tom Mitchell. Machine Learning, chapter Decision Tree Learning, pages 52–78. McGraw-Hil, 1997.
- Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey. Phishing websites features,

2015. Unpublished. Available via: http://eprints. hud. ac. uk/24330/6/RamiPhishing_ Websites_Features. pdf.

- Rami M. Mohammad, Fadi A. Thabtah, and Lee McCluskey. An assessment of features related to phishing websites using an automated technique. In 7th International Conference for Internet Technology and Secured Transactions, ICITST 2012, London, United Kingdom, December 10-12, 2012, pages 492–497, 2012.
- S´anchez-Maro no, Noelia, Alonso-Betanzos, Amparo, and Mar´?a Tombilla-Sanrom´an. Filter methods for feature selection: A comparative study. In Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL'07, pages 178–187, 2007.
- Jasmina Novakovic, Perica Strbac, and Dusan Bulatovic. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1):119–135, 2011.
- Selwyn Piramuthu. Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research, 156(2):483–494, 2004.
- M. Ramaswami and R. Bhaskaran. A study on feature selection techniques in educational data mining. CoRR, abs/0912. 3924, 2009.
- Yvan Saeys, I˜naki Inza, and Pedro Larra˜naga. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19):2507–2517, September 2007.
- Luis Talavera. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In 6th International Symposium on Intelligent Data Analysis, IDA'05, 2005.
- Jason Weston. Support vector machine tutorial.
- Joshua S. White, Jeanna N. Matthews, and John L. Stacy. A method for the automated detection phishing websites through both site characteristics and image analysis, 2012.
- Colin Whittaker, Brian Ryner, and Marria Nazif. Large-scale automatic classification of phishing pages. In Proceedings of the Network and Distributed System Security Symposium, NDSS 2010, San Diego, California, USA, 28th February - 3rd March 2010, 2010.
- Guang Xiang, Jason I. Hong, Carolyn Penstein Ros´e, and Lorrie Faith Cranor. CANTINA+: A feature-rich machine learning framework for detecting phishing web sites. ACM Trans. Inf. Syst. Secur. , 14(2):21, 2011.
- Yue Zhang, Jason I. Hong, and Lorrie Faith Cranor. Cantina: a content-based approach to detecting phishing web sites. In Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007, pages 639–648, 2007.

Index Terms

Computer Science Information Sciences

Keywords

phishing websites detection machine learning classification feature selection