基于改进随机森林算法的钓鱼网站检测方法研究

朱　琪　　1,2　,林果园　　1,2,3
(1 中国矿业大学 计算机科学与技术学院，江苏 徐州 221116；2 矿山数字化教育部工程研究中心，
江苏 徐州 221116；3 南京大学 计算机软件新技术国家重点实验室，江苏 南京 210023)
摘　要：为了更准确快捷的对钓鱼网站进行识别，提出了一种基于改进随机森林算法的钓鱼网站检测方法.该方法挖掘钓鱼网页特征之间潜在的关联规则，并对数据集进行分区，以此区分特征数据的重要程度并计算权重以及数据选取的比例，选取数据后对数据空间进行相应的集合化与剪辑以此优化森林的建立，并根据建立的森林达到对钓鱼网站检测识别的目的.最终实验说明，该方法对钓鱼网站的检测识别具有很好的效果和效率.
关键词：钓鱼检测；关联规则；特征分区；数据空间

Research on Detection Methods of Phishing Websites Based on Improved Random Forest Algorithm

ZHU Qi　　1,2　, LIN Guo-yuan　　1,2,3
(1 School of computer science and technology, China University of Mining and Technology, XuZhou 221116, China;
2 Mine Digitization Engineering Research Center of the Ministry of Education, XuZhou 221116, China;
3 State Key Laboratory for Novel Software Technology，Nanjing University,NanJing 210023 , China)
Abstract：In order to improve the efficiency of phishing detection，a new algorithm was proposed to improve the traditional random forest algorithm. Potential association rules between web features are mined and used to partition the data set,in order to distinguish the features of different structures and calculate the weight of different data space to determine the scale of the selection. After selection of data, training data sets need to be aggregated and clipped to optimize the establishment of forests. Websites are trained and predicted using voting in decision forest. Experiments result shows that the new algorithm has obvious advantages in efficiency and effectiveness compared with the other two algorithm.
Key words：fishing detection; association rules; feature partition; data space
作者简介：
朱　琪　男，(1994-)，硕士研究生.研究方向为云计算与信息安全.E-mail:747116218@qq.com.
林果园　男，(1975-)，博士，副教授.研究方向为网络空间安全、移动互联及其安全、云计算及其安全、信息系统及其安全.