


# A Study of Feature Selection and Dimensionality Reduction Methods for Classification-Based Phishing Detection System

Amit Singh, Indian Computer Emergency Response Team, India

Abhishek Tiwari, Central University of Haryana, India

 <https://orcid.org/0000-0002-6549-8132>

## ABSTRACT

Phishing was introduced in 1996, and now phishing is the biggest cybercrime challenge. Phishing is an abstract way to deceive users over the internet. Purpose of phishers is to extract the sensitive information of the user. Researchers have been working on solutions of phishing problem, but the parallel evolution of cybercrime techniques have made it a tough nut to crack. Recently, machine learning-based solutions are widely adopted to tackle the menace of phishing. This survey paper studies various feature selection method and dimensionality reduction methods and sees how they perform with machine learning-based classifier. The selection of features is vital for developing a good performance machine learning model. This work is comparing three broad categories of feature selection methods, namely filter, wrapper, and embedded feature selection methods, to reduce the dimensionality of data. The effectiveness of these methods has been assessed on several machine learning classifiers using k-fold cross-validation score, accuracy, precision, recall, and time.

## KEYWORDS

Cyber-Crime, Dimensionality Reduction, Feature Selection, Machine Learning, Phishing

## 1. INTRODUCTION

In Phishing, the phisher creates a fraud phishing website to mislead web users to steal their sensitive personal information. Deception is the way of Phishing by hiding as a trusted entity in electronic communication. The first time Phishing discovered in the 1980s. Anti-Phishing Working Group (APWG) reported 51,401 unique phishing websites in June 2018 (Chiew, Tan, Wong, Yong, & Tiong, 2019; *Phishing Activity Trends Report 2nd Quarter 2018*, 2018). Another report by RSA estimated that global organizations lost 9 billion\$ due to phishing fraud in 2016 (Heidi Bleau, 2016). It is one of the biggest cybercrime faced by internet users. Generally, phishing attacks are accomplished using emails and website spoofing. Phishers start the attack by sending spoofed emails to victims and

DOI: 10.4018/IJIRR.2021010101

victims think this is authentic and secure, thereby they got trapped. Figure 1 represents the workflow structure of phishing.

Apart from email, phisher leads users to various similar looking authenticated, secure and famous websites via advertisement links. There are many ways of phishing detection and prevention such as the use of any authorized anti-phishing software, naive browser extensions (Google and Mozilla Firefox use Blacklist warning system) and toolbars. Blacklist warning system queries a database of already known phishing URLs so it will not be able to identify new upcoming phishing websites (Chiew et al., 2019). Designing an intelligent phishing detection system, based on Machine learning classification model can easily identify whether this website or web-link is for phishing or not. These ML based classification systems are very effective. However, for creating these prediction system in machine learning, feature selection and dimensionality reduction are very important steps. Investigation of state of the art approaches reveals that there is a need for a systematic study of feature selection and dimensionality reduction approaches to design an intelligent and capable system to detect the phishing websites.

Figure 1. Phishing workflow



For any Machine learning classifier, we need useful and relevant features. For choosing, those relevant features from the dataset feature selection is paramount. Feature selection is even more useful when we are dealing with high dimensional data. This high dimensional dataset poses many problems, such as increased training time and sometimes it may lead towards overfitting of our machine-learning model. The feature selection process will select relevant attributes from data based on the method specified by the analyst (Ameen, Balogun, Usman, & Fashoto, 2016). These reduced features will help us in improving the accuracy of the classifier and decrease the computational cost of the classifier. There are three main category of feature selection techniques filter method, wrapper method, and embedded method. All these techniques have their unique significance, and we will discuss it section 3.

Dimension reduction is another feature preprocessing technique before the design of a classifier. Dimensional reduction transforms the dataset into a low dimensional dataset, ensuring it will not change the meaning of data. When the dimensionality of the datasets reduced, then it improves the performance of the classifier in comparison to applying on original data. Dimensionality reduction can be both linear and nonlinear; it depends on the dataset.

Feature selection and Dimensionality reduction both are used in designing the best Machine learning Classification model with a difference that features selection technique aims at selecting the features from original dataset whereas dimensionality reduction technique aims at transforming the dimensionality of original datasets.

Machine learning focuses on developing the computation algorithms to find out patterns, reasoning, and rules from data to design Machine Learning model, which can detect or make a prediction about forthcoming occurrences (Ali, 2017). Machine learning is supervised learning if outputs are given with training data for training the model else; it is unsupervised learning. Many supervised learning algorithms are successfully working on real-life applications. Some popular Machine learning Classification techniques are Support Vector machine (SVM), Naïve Bayes classifier, K Nearest Neighbor (KNN), Decision trees, Random forest, and Ensemble methods. These

33 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/a-study-of-feature-selection-and-dimensionality-reduction-methods-for-classification-based-phishing-detection-system/267120](http://www.igi-global.com/article/a-study-of-feature-selection-and-dimensionality-reduction-methods-for-classification-based-phishing-detection-system/267120)

## Related Content

---

### A Highest Sense Count Based Method for Disambiguation of Web Queries for Hindi Language Web Information Retrieval

Sanjay K. Dwivedi (2012). *International Journal of Information Retrieval Research* (pp. 1-11).

[www.irma-international.org/article/a-highest-sense-count-based-method-for-disambiguation-of-web-queries-for-hindi-language-web-information-retrieval/90438](http://www.irma-international.org/article/a-highest-sense-count-based-method-for-disambiguation-of-web-queries-for-hindi-language-web-information-retrieval/90438)

### Extracting Entities of Emergent Events from Social Streams Based on a Data-Cluster Slicing Approach for Ontology Engineering

Chung-Hong Lee and Chih-Hung Wu (2015). *International Journal of Information Retrieval Research* (pp. 1-18).

[www.irma-international.org/article/extracting-entities-of-emergent-events-from-social-streams-based-on-a-data-cluster-slicing-approach-for-ontology-engineering/132499](http://www.irma-international.org/article/extracting-entities-of-emergent-events-from-social-streams-based-on-a-data-cluster-slicing-approach-for-ontology-engineering/132499)

### An Enhanced Multi-Frequency Distorted Born Iterative Method for Ultrasound Tomography Based on Fundamental Tone and Overtones

Tran Quang-Huy, Tuan-Khai Nguyen, Vijender Kumar Solanki and Duc-Tan Tran (2022). *International Journal of Information Retrieval Research* (pp. 1-19).

[www.irma-international.org/article/an-enhanced-multi-frequency-distorted-born-iterative-method-for-ultrasound-tomography-based-on-fundamental-tone-and-overtones/289608](http://www.irma-international.org/article/an-enhanced-multi-frequency-distorted-born-iterative-method-for-ultrasound-tomography-based-on-fundamental-tone-and-overtones/289608)

### Experiments and their Assessment

Ibrahim Dweib and Joan Lu (2013). *Design, Performance, and Analysis of Innovative Information Retrieval* (pp. 249-263).

[www.irma-international.org/chapter/experiments-their-assessment/69141](http://www.irma-international.org/chapter/experiments-their-assessment/69141)

### An Efficient Innovative Approach Towards Color Image Enhancement

Dibya Jyoti Bora (2018). *International Journal of Information Retrieval Research* (pp. 20-37).

[www.irma-international.org/article/an-efficient-innovative-approach-towards-color-image-enhancement/193247](http://www.irma-international.org/article/an-efficient-innovative-approach-towards-color-image-enhancement/193247)