Heuristic nonlinear regression strategy for detecting phishing websites

Article *in* Soft Computing · June 2019

DOI: 10.1007/s00500-018-3084-2

3 authors, including:

Mehdi Babagoli
Khaje Nasir Toosi University of Technology
3 PUBLICATIONS **80** CITATIONS

SEE PROFILE

Vahid Solouk
Urmia University of Technology
37 PUBLICATIONS **261** CITATIONS

SEE PROFILE

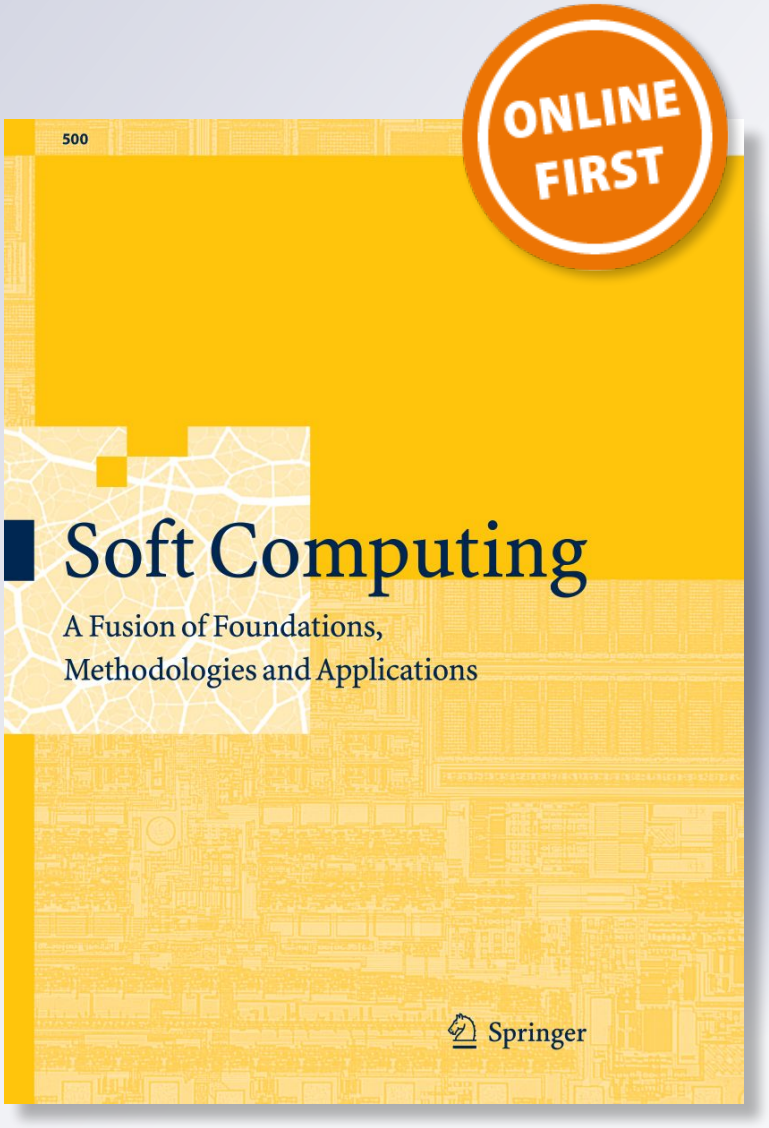
Heuristic nonlinear regression strategy for detecting phishing websites

Mehdi Babagoli, Mohammad Pourmahmood Aghababa & Vahid Solouk

Soft Computing
A Fusion of Foundations, Methodologies and Applications

ISSN 1432-7643

Soft Comput
DOI 10.1007/s00500-018-3084-2



Springer

Soft Computing
https://doi.org/10.1007/s00500-018-3084-2

METHODOLOGIES AND APPLICATION

Heuristic nonlinear regression strategy for detecting phishing websites

Mehdi Babagoli[1] · Mohammad Pourmahmood Aghababa[2] · Vahid Solouk[1]

Abstract
In this paper, we propose a method of phishing website detection that utilizes a meta-heuristic-based nonlinear regression algorithm together with a feature selection approach. In order to validate the proposed method, we used a dataset comprised of 11055 phishing and legitimate webpages, and select 20 features to be extracted from the mentioned websites. This research utilizes two feature selection methods: decision tree and wrapper to select the best feature subset, while the latter incurred the detection accuracy rate as high as 96.32%. After the feature selection process, two meta-heuristic algorithms are successfully implemented to predict and detect the fraudulent websites: harmony search (HS) which was deployed based on nonlinear regression technique and support vector machine (SVM). The nonlinear regression approach was used to classify the websites, where the parameters of the proposed regression model were obtained using HS algorithm. The proposed HS algorithm uses dynamic pitch adjustment rate and generated new harmony. The nonlinear regression based on HS led to accuracy rates of 94.13 and 92.80% for train and test processes, respectively. As a result, the study finds that the nonlinear regression-based HS results in better performance compared to SVM.

Keywords Phishing · SVM · Harmony search · Feature selection · Decision tree · Wrapper · Nonlinear regression

1 Introduction

As a fundamental component of the daily social activities, the Internet is in a ubiquity proliferation apart from the developers' main goals, and due to the users are constantly exposed to online threats. Such threats may lead to the compromise of some important financial and personal data losses and identity in e-commerce (Mohammad et al. 2014b). Among various types of threats, phishing is referred to as deception in e-commerce with attempts to steal confidential information of the users through impersonating the target website. Typically, in a phishing threat, the photographs and contents of the fraudulent websites are similar to the legal websites (Basnet et al. 2008; Gupta and Shukla 2015). On the other hand, finding a solution to identify all phishing websites poses specific challenge due to considering the complexity of phishing procedure and developing ways regarding these attacks. Generally, the phishing detection methodologies can be divided into two categories: intelligent and traditional schemes. The intelligent methods (such as genetic algorithm, particle swarm optimization, harmony search, ant colony optimization, etc.) are eventually inspired from natural phenomena with decision-making ability (Mohammad et al. 2014b). While the heart of decision-making process in such intelligent algorithms are based upon training with some suitable data, the traditional approaches (such as rule-base methods, white-list, black-list, hash-list and extended black-list) require no training. In addition, traditional algorithms operate implicitly and require no classification, which leads to less execution time.

Modern browsers such as Firefox and Netcraft use generally use black-list databases, i.e., a comprehensive list of fraudulent websites in order to deal with phishing attacks. Accordingly, when a URL is requested through the browser, the system queries the database for the URL and if the entry exists, the webpage is blocked. Such methods might deem as inadequate-in-sole solutions, because the phishers can pass through some filters using fake addresses. As a result, improvements in those traditional methods realize through integrations with other solutions to decrease the risk

Communicated by V. Loia.

✉ Mohammad Pourmahmood Aghababa
m.p.aghababa@ee.uut.ac.ir; m.p.aghababa@gmail.com

[1] Faculty of Computer Engineering, Urmia University of Technology, Urmia, Iran

[2] Faculty of Electrical Engineering, Urmia University of Technology, Urmia, Iran

Published online: 20 February 2018

of vulnerabilities (Abdelhamid et al. 2014). There are number of studies (Aburrous et al. 2008; Gupta and Shukla 2015; Mohammad et al. 2014b) conducted to introduce methods based on using features for identifying legitimate websites from those fraudulent. The feature are further used as the basic knowledge of meta-heuristic algorithms or neural networks (Aburrous et al. 2010). Some of the features include an IP Address within the URL, spelling error, and abnormal DNS record.

Meta-heuristic algorithms are higher-level procedures which are designed to find, generate, or select a heuristic (partial search algorithm) that may provide a good solution for an optimization problem. Over the past five decades, many algorithms have been developed to solve engineering optimization problems. Most of the developed algorithms are based on linear or nonlinear programming approaches. However, there are some complex problems with no solutions using either a linear or a nonlinear programming method. For instance, if the problem contains more than one local optimal solution, the pertaining method must start with different initial points. Meta-heuristic alternatives are able to find optimal solution in complex problems using their capabilities (combination of randomness and rules, high speed, etc.) (Lee and Geem 2005). General classification of meta-heuristics is shown in Fig. 1 based on their operational procedure. Some of the procedures have used a dataset to classify the features which are effective in the phishing detection (Hamid and Abawajy 2011; Mohammad et al. 2012; Montazer and ArabYarmohammadi 2013), and some other procedures have proposed heuristic algorithms to detect the phishing websites. One of the best solutions to detect fraud websites is the identification of the websites' properties and modeling phishing websites based on their characteristic. According to this issue, there are various methods for modeling of the dynamic systems (Qiu et al. 2017; Wei et al. 2017). The phishing websites can be modeled by their properties which could lead to reduce the computational cost.

In order to improve the accuracy and the efficiency of the phishing detection mechanism, the current paper proposes a detection solution based on a nonlinear regression method. In the study, we used a dataset from the UCI Database (Mohammad et al. 2015). The dataset consists of 11055 website instances (rows) and 31 features (columns). We used two feature selection methods, namely decision tree (DT) and the wrapper. The feature selection techniques were utilized to remove the irrelevant attributes and to reduce the train time. After feature selection, a model of nonlinear regression (NR) is suggested and then a modified harmony search (MHS) is used to find the optimal parameters of the proposed model. The nonlinear regression based on harmony search (NR-MHS) and support vector machine (SVM) are used to predict the fraudulent websites. This research shows that using of meta-heuristic algorithms confirms better performance in comparison with some other heuristic algorithms.

The rest of this paper is organized as follows. Section 2 is devoted to a literature review regarding the previous works. In Sect. 3, the proposed phishing detection method is presented. The experimental results and discussion are shown in Sect. 4. At last, Sect. 5 ends this paper with conclusions.

2 Literature review

In this section, some related studies about phishing detection are reviewed. Mohammad et al. (2014b) have used the artificial neural network to detect phishing websites. The applied neural network consists of 17 input neurons that show the number of the selected features. Their work indicated that the hidden layer can include one or more neurons. Furthermore, 80% of data has been used for train and 20% of data has been adopted for test. The testing accuracy of the prediction has been obtained 92.48% in 500 epochs. Hamid and Abawajy (2011) have used hybrid-feature selection method to detect the phishing E-mails. Seven features have been used to predict fraudulent websites, and the detection accuracy of about 93% has been obtained. Montazer and ArabYarmohammadi (2013) have prepared some questionnaires to access expert's view point about the degree of importance of each features in Iranian's e-banking. In their research, 40% of questionnaires have been returned and the results have been averaged. After gathering respondent data, they have used the exploratory factor analysis to determine the critical indicators which were effective on phishing detection in Iranian e-banking system. The average value of features has been divided into the same range between 5 and 8, which means the "Medium" and the "Much" importance. Some features have been selected as more important factors among all of 28 features. The selected features were: the server form handler (SFH), distinguished names certificate (DN), disabling right click, using hexadecimal character codes and abnormal cookie. In Pandey and Ravi (2012), data and text mining methods have been applied to detect the phishing E-mails. The dataset used in Pandey and Ravi (2012) consists of 2500 phishing and legitimate E-mails. The text mining has been used to select 23 features from email body. Then, the t-static method has been used to choose the most important features. They have used the multi-layer perceptron (MLP), decision tree, SVM, group method of data handling, genetic programming and logistic regression for classification. As shown in their results, the MLP confirms a better accuracy than the other methods. The accuracy has been obtained 98.12% for MLP. According to the prevalence of social media network like twitter, Jeong et al. (2016) have used a 2-phase clustering algorithm which is called PDT (phishing detector for twitter) to detect the phishers, scammers and spammers. The features which

Fig. 1 General classification of meta-heuristic method

have been adopted in this research have been divided into the three groups: tweet features, user features and URL features. They have obtained a variable accuracy for phishing detection (between 0.88 and 0.99). The variable accuracy is the most important weakness of this approach. Forwarding-base features were used in Cao et al. (2016) to detect the malicious URLs in online social networks. The authors of Cao et al. (2016) have used the Bayes net, J48 and random forest to detect the phishing URLs. As shown in their results, the average accuracy reached to 83.21%.

However, most of the above-mentioned methods suffer from some restrictions which include: lack of stability to change within phishing tricks, disability to detect the phishing websites with constant accuracy, the lack of comprehensive dataset, difficulty in recognizing phishers before they

are attracted the users and inefficiency of list-based methods for new phishing websites.

3 Proposed phishing detection method

Hypothetically, websites are presumed to contain plenty of information, from which reasonable sets of features can be extracted (Aburrous et al. 2010; Mohammad et al. 2014b; Montazer and ArabYarmohammadi 2013). As a side effect, excess number of features can lead to inaccurate decisions due to deterioration of the resources and thereby, degradation of detection performance. For example, the required CPU time (runtime) can relatively increase by increasing the number of features (Wang et al. 2014). The features are analyzed and evaluated with the DT and wrapper approaches. Finally, in the proposed phishing detection algorithm, a modified HS and the SVM techniques are used to detect and predict the phishing websites.

3.1 Phishing dataset

The phishing dataset used in this research is adopted from the UCI Datasets (Mohammad et al. 2015) and is comprised of 31 columns and 11055 rows, consisting of 30 features (see Table 1) with the value of each feature being − 1 (Phishing), 0 (Suspicious) or 1 (Legitimate). The last column of the dataset includes the results of each sample, with phishing denoted using the value − 1 and legitimate denoted using the value 1. Hence, each row represents a legitimate or a phishing website. The detailed description of the features can be found in Mohammad et al. (2012, 2014a, b).

3.2 Feature selection method

Prior to our phishing detection approach being employed, DT and wrapper methods are applied in two phases to achieve a clear penetration of the feature set and remove the noisy features from the dataset. DT is applied in the first phase. In this approach, once the elimination of the nodes in the sub-tree does not affect the root, the feature located in the root considered as an important feature (Fig. 2a, b). When the most important feature is found, it is removed from the DT list and the next important feature will be replaced in the root (Fig. 2c, d). This procedure continues until the accuracy of the DT is decreased significantly.

In the second phase, the wrapper procedure with genetic algorithm (GA) search method is implemented to select the best feature subset (Rodrigues et al. 2014). The classification algorithms within the wrapper methods are considered as a black box. Therefore, the classification methods are used as an evaluator for the feature subset selection and the heuristic search methods are employed to find the optimal subsets for the classification methods (Song et al. 2017). The wrapper method of feature selection is performed with GA using the DT classifier as a black box. In the wrapper method, the original features are embedded into the GA algorithm applied to find the optimal feature subset with the high train accuracy which is gained by the DT.

Initially, the dataset is divided into 10 segments (folds) wherein the GA selects 9 fold as training sequence and 1 fold as test. At each iteration, the accuracy of the selected segments (fold) is evaluated through the DT. The procedure continues until the best sets are chosen for training. Figure 3 shows the procedure of the proposed wrapper method. The wrapper method of feature selection is implemented in Waikato Environment for Knowledge Analysis (WEKA) (Hall et al. 2009).

3.3 Proposed regression model

In this paper, we propose a nonlinear regression (NR) based on HS to detect the phishing websites using the extracted feature. The nonlinear regression attempts to find the functional relationship between the inputs and outputs(Fil et al. 2016). Here, the coefficients of the nonlinear regression are estimated by a modified HS (MHS). The proposed MHS is designed to minimize the mean-square-error (MSE) between the predicted and target outputs. The following model is used in the proposed approach as the cost function.

$$F(r) = \text{Sign}\left(\sum_{i=1}^{N} \alpha_i x_i + \sum_{t=1}^{N} \sum_{j=t+1}^{N} \alpha_{tj} x_t x_j + \beta\right) \tag{1}$$

where, r denotes the row, N shows the number of selected features, α is a harmony, β is a random number between $[-1, 1]$ and x denotes the input vector which shows the instances of websites and includes 20 features. The sign function of a real number x is defined as follows.

$$sign(\text{x}) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{2}$$

Finally, MSE is calculated for each row (vector) of dataset matrix as follows.

$$\text{MSE} = \frac{\sum_{r=1}^{M} (F(r) - F(s))^2}{M} \tag{3}$$

where $F(s)$ represents the desired output, M denotes the number of rows and $F(r)$ is obtained from Eq. (1).

Table 1 The features of the UCI dataset

1.	*Having_IP_Address*: Using IP address instead of domain name in websites can show that somebody wants to steal the information
2.	*URL_Length*: Phishers try to hide the suspicious part of URL by using the long URL
3.	*Shorteting_Service*: using this method leads to the URL which may be made significantly shorter in length. Short URLs can be directed to the phishing websites
4.	*Having_At_symbol*: The browsers disregard everything which is typed before "@" symbol. Then, phishers put their fake address after "@" symbol
5.	*Double_slash_redirecting*: The "//" symbol in URL is used to redirect the users to another page. If the number of "//" in the URL is more than 7, the website is obtained as phishing
6.	*Prefix_Suffix*: Prefix or suffix are added to the URL with dash symbol. By using this technique, users can not be able to detect the difference between legitimate websites and phishing websites
7.	*Having_Sub_Domain*: Sub domains are separated by dot symbol in URL. If the number of sub domains are more than 1, the website is obtained as fraudulent website
8.	*SSLfinal_state*: The content of certificate which are used in the websites must have a trusted issuer (GeoTrust, GoDaddy, Thawte and etc.) and the age of the certificate must be more than 1 year. If a HTTPS website has the mentioned characteristic, it can be regarded as legitimate website
9.	*Domain_registration_length*: Phishing website lives for a short time. Typically, if the registration length is less than one years, the website is regarded as phishing
10.	*Favicon*: Favicon is an icon which is represented in the address bar with URL. In the phishing websites, favicons are loaded from the URL which is not the same as the domain name in the address bar
11.	*Port*: Standard URL must use a specific port number (80 or 443). The phisher use non-standard port number which aims to steal user's information
12.	*HTTPS_token*: existence of "HTTPS" in domain part of URL is the symptom of phishing websites
13.	*Request_URL*: In the phishing websites, the objects are loaded from different domains. If more than 66% of objects are loaded from various domains, the feature is regarded as fraudulent
14.	*URL_of_Anchor:* links in the websites are placed in <a> tags. If more than 61% of the anchor tags are irrelevant to the webpage name, the feature is defined as phishing
15.	*Links_in_tags*: the phishers are deceive people by using the fake address. The fake address could be placed in <link>, <meta> and <script> tags.
16.	*SFH*: Forms in the websites are managed with server form handler (SFH). When the users fill the form and submit their information, SFH is obliged to show a message. In phishing websites, the SFHs are contained an empty string," about: blank" message or fraudulent URL address
17.	*Submitting_to_email:* If server form handler use "mail ()" or "mailto ()" functions, the feature is regarded as phishing
18.	*Abnormal_URL*: legitimate websites have used the hostname for an identity of their URLs. Normal URLs must include the name of the site
19.	*Redirect*: The phishers use consecutive redirects to confuse the users
20.	*On_mouseover*: the address which is shown on the status bar must be same as to the URL in the address bar. This feature checks these URLs and if they are different with each other the feature is regarded as phishing
21.	*RightClick*: The phishing websites hide their source code by disabling the right click
22.	*popUpWindow*: The legitimate websites use a specific form to get the personal information. But in the phishing websites, the popup window is used to save the personal data
23.	*Iframe*: Phishers use <Iframe> tags to display another webpage in the current website
24.	*Age_of_domain*: Most of the phishing websites live for a few months. After a few months, phishers destroy the website contents
25.	*DNSRecord*: For phishing websites, the DNS record is not recognized in WHOIS datasets (Mohammad et al. 2015)
26.	*Web_traffic*: This feature represents the total visit of websites and their importance. The importance of websites is demonstrated by their ranking in the Alexa website (www.alexa.com)
27.	*Page_Rank*: Page rank shows the importance of websites and it is valued between (0, 1). Phishing websites are placed in (0, 0. 2)
28.	*Google_index*: The legitimate websites are indexed by Google. If a website does not exist into the Google list, the feature is labeled as phishing
29.	*Links_pointing_to_page*: The number of links pointing to the page shows the legitimacy level of websites. The phishing websites have no links pointing to them
30.	*Statistical_report*: Some of the organization works on the phishing websites. They detect the fraudulent websites and add them to their database. The statistical report uses these organization databases

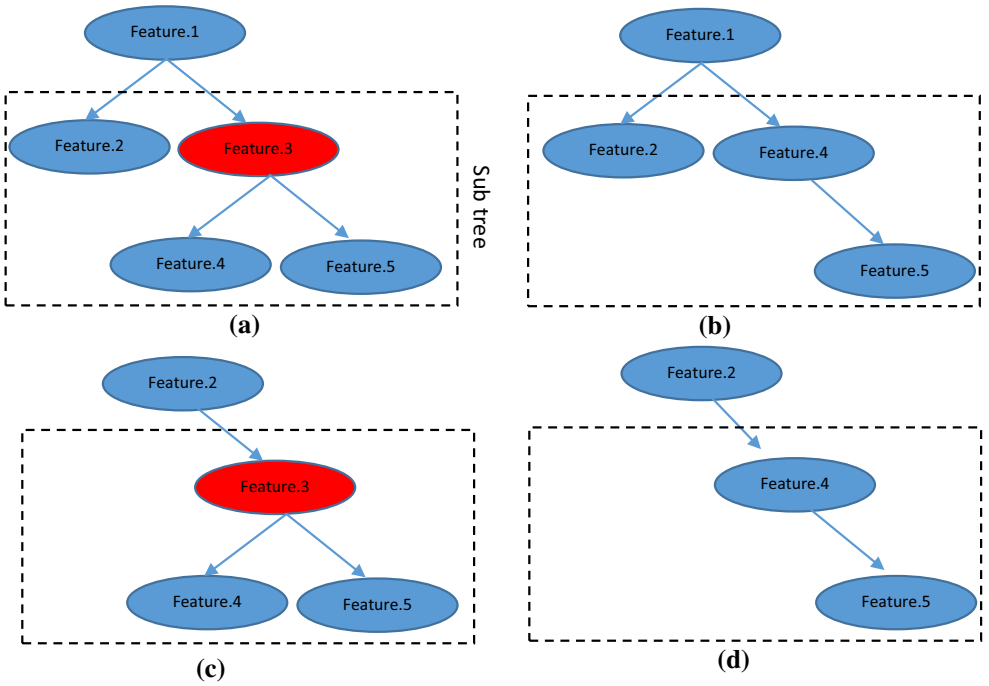
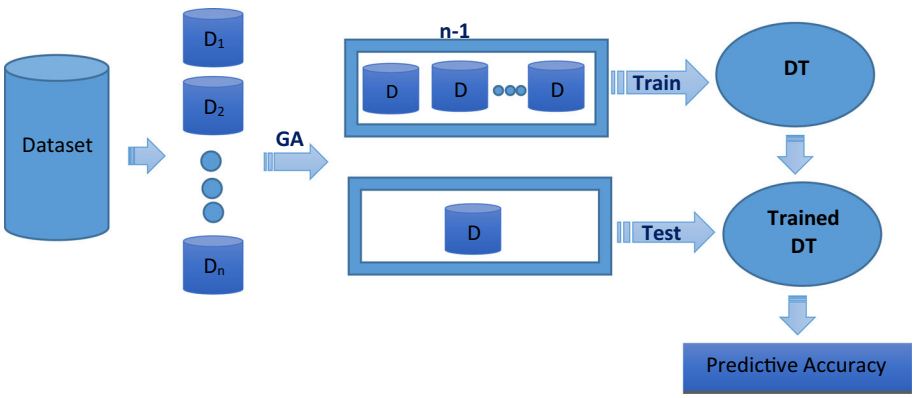


Fig. 2 Feature selection with DT method. **a** Decision tree. **b** Eliminate the feature.3. **c** Eliminate the feature.1. **d** Eliminate the feature.3

Fig. 3 Feature selection with wrapper method



3.4 Main procedure of the proposed phishing detection approach

In this article, the nonlinear regression based on modified harmony search and the SVM classification are used for phishing detection. The methods are described in details as below.

3.4.1 Nonlinear regression based on harmony search

As mentioned earlier, the nonlinear regression is a regression analysis that uses a combination of the independent variables to solve the nonlinear problems. Most of the researches use optimization algorithms and neural networks to achieve the best weights for the NR model (He et al. 2016; Satapathy et al. 2012). In this study, the harmony search is used to estimate the best weights for the NR. Harmony search method is a meta-heuristic algorithm which is used for optimization problems. HS is inspired from the process of musical performances (Ameli et al. 2016; Wang et al. 2016). In this algorithm, a solution vector is similar to a harmony in music and searching for solution vector is the same as the process used by an orchestra (looking for the best harmony among all available modes for playing) (Manjarres et al. 2013). The advantages of HS in comparison with the other meta-heuristic algorithms are using the stochastic search based on the pitch adjustment rate and the harmony memory consideration rate (Kalivarapu et al. 2016). Figure 4 illustrates the modified harmony search flowchart. In order to increase the accuracy of the traditional HS and to give the ability of escaping from local optima, a modified HS for phishing detection is proposed in this paper as below.

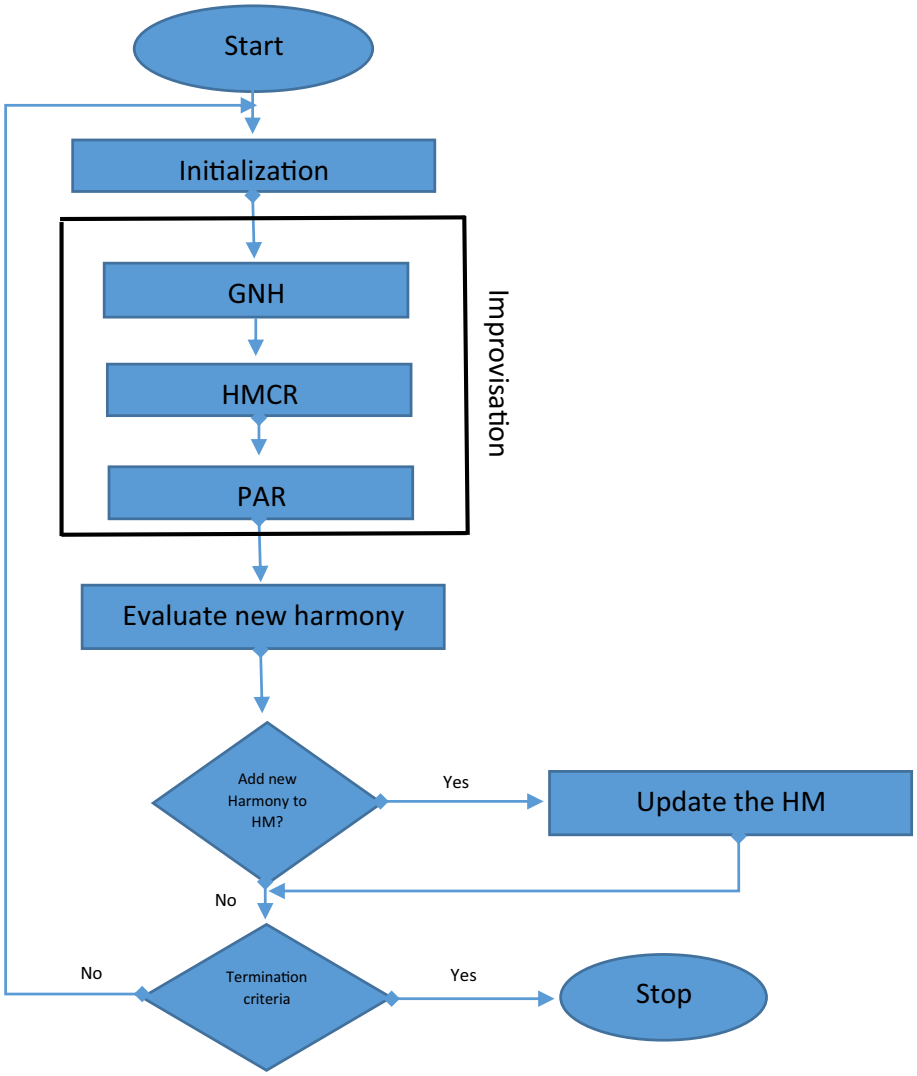


Fig. 4 Flowchart of the proposed modified harmony search

Step 1 Initialize the default parameter of HS: HMCR, PAR, HMS and BW.
where HMCR is a probability of new harmony selection from harmony memory, PAR presents the probability of the new harmony obtained by adding a small random value between $[-1, 1]$, HMS and BW are the size of harmony memory (in this work obtained 30) and the bandwidth of decision value (between $[-1, 1]$), respectively.
Step 2 Initialize harmony memory (HM) by a random matrix of containing values in range $[-1, 1]$.

$$\mathrm{HM} = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,M} \\ \vdots & \ddots & \vdots \\ \alpha_{\mathrm{HMS},1} & \cdots & \alpha_{\mathrm{HMS},M} \end{pmatrix} \tag{4}$$

Step 3 Generate a new harmony (α'_{new}) vector. The α'_{new} can be chosen from HM with the HMCR probability: $\alpha'_{new,i} \in \{\alpha_{1,i}, \alpha_{2,i}, \ldots, \alpha_{\mathrm{HMS},i}\}$, and with the 1-HMCR probability, it can be equal to a random number between $[-1, 1]$. If the new harmony is chosen from HM, with PAR probability, the α'_{new} will be summed with a random number (DELTA) between $[-1, 1]$ ($\alpha'_{\mathrm{new}} = \alpha'_{\mathrm{new}} +$ DELTA).In the common harmony search, the decision is made separately for each element of new harmony and the number of selected elements in the new harmony is constant (equal to 1) but in the proposed harmony search the number of the new generated harmony elements can be changed between 1, 3, 5 and 7, in each iteration. Figure 5 illustrates an example of generating new harmony, where the number of the generated new harmony (GNH) is set to 3.
The PAR is also changed dynamically in each iteration (Naik et al. 2016).

$$\mathrm{PAR} = \frac{(\mathrm{PAR}_{\max} - \mathrm{PAR}_{\min})}{(\max\ \mathrm{Itteration}) \times \mathrm{current\,Itteration} + \mathrm{PAR}_{\min}} \tag{5}$$

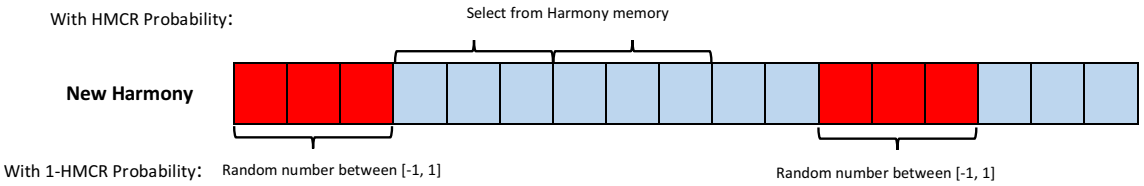


Fig. 5 New harmony generation with GNH = 3

Step 4 Replace the worst vector in the HM by the new vector, if the new vector is better than the worst one.
Step 5 Repeat Steps 2–4 until a termination criterion is obtained.

3.4.2 Support vector machine (SVM)

Support vector machine (SVM) is a supervised learning method that analyzes the data used for classification and regression (Cai et al. 2003). In linear separable cases, SVM constructs a hyperplane to separate two different classes. The hyperplane is constructed by finding vector w and parameter b that minimizes $\|w\|^2$ and satisfies the following conditions considering the training data as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

$$\text{Minimize} : \frac{1}{2} w^T \cdot w \tag{6}$$

$$\text{Subject to} : y_i(w^T \cdot x_i + b) \geq 1 \tag{7}$$

where w is the weight vector, x is the input vector, y is the classes label and b represents the bias term. To deal with cases where there may be not separable due to noisy data, the soft margin SVM is proposed in Xia et al. (2016). The SVM changes into the following model when the case consists of non-separable data due to some noises.

$$\text{Minimized} : \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{N} \xi_i \tag{8}$$

$$\text{Subject to} : y_i(w^T \cdot x_i + b) \geq 1 - \xi_i, i = 1, 2, 3, \ldots, n \tag{9}$$

where $C \geq 0$ is a parameter that controls the amount of training error and ξ_is represents the nonnegative slack variables which are misclassified. In this work the amount of C is chosen based on trial and error method. The solution procedure indicates that the best value for this parameter is 1.

Remark 1 According to above equations, x and y are feature vectors and classes, respectively. On the first step, LibSVM tries to find the optimal w vector which must satisfy the main condition (Eq. 8) and after that, w is substituted in Eq. 9 to obtain the test and train accuracies. Karush–Kuhn–Tucker (KKT) conditions are the first-order requirements for a solution to the nonlinear convex optimization problem(Jahn 2017). KTT can be investigated to guaranty the feasibility of the proposed algorithm.

Table 2 Description of the dataset (Mohammad et al. 2012)

Dataset	Cases	Classes	Attributes
Phishing websites	2456	2	30

4 Simulation results

In order to evaluate the performance of our approach, we conducted simulation analyses as described in this part. The feature selection and phishing detection methods were implemented in Weka3. 6. 0 and MatlabR2014a, respectively. All the simulation runs were implemented on 2.00 GHz processor with 6 GB of random access memory.

4.1 Results of NR-MHS and SVM

Both SVM and HS methods were initially structured with a random population in the range of $(-1, 1)$. The benchmark dataset which was used in this paper selected from UCI database, which consists of 11055 rows and 31 columns and three references (Mohammad et al. 2012, 2014a, b). Table 2 lists partial details of the dataset.

In the feature selection, two methods including the wrapper method (with genetic search method and 10-fold cross-validation) and the DT algorithm are compared with each other. As seen in Figs. 6 and 7, the 20 most important terms are chosen among 31 features.

Figure 6 depicts the features which have been selected using the DT method. Each value on top of the bars represents the accuracy of the decision tree when the specific feature is chosen as the root. For example, "SSL_finalstate" is the most important feature for two reasons: (1) it is placed in the root when the DT is plotted for all features, and (2) the elimination of the other features does not significantly affect the accuracy of the DT. As shown in Fig. 7, the blue bars represent the selected features and the red bars show unselected features. The number on the top of the bar shows the merit

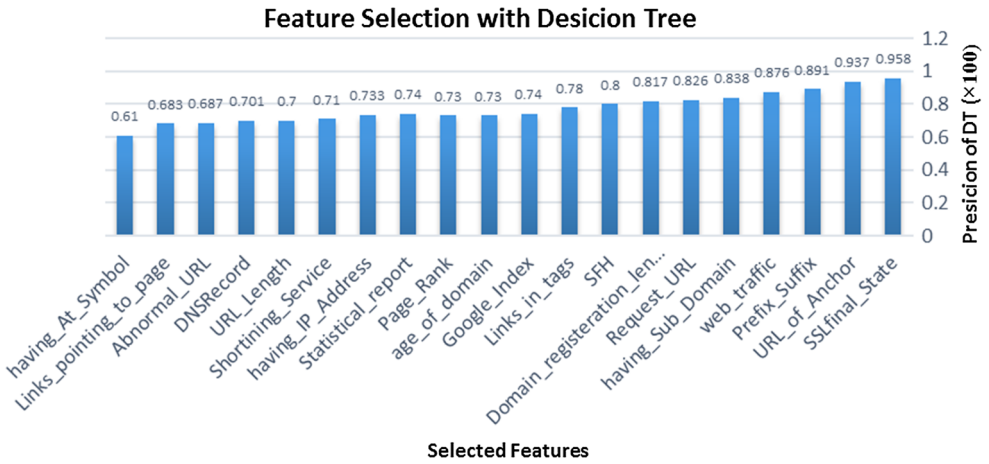


Fig. 6 Feature selection with decision tree

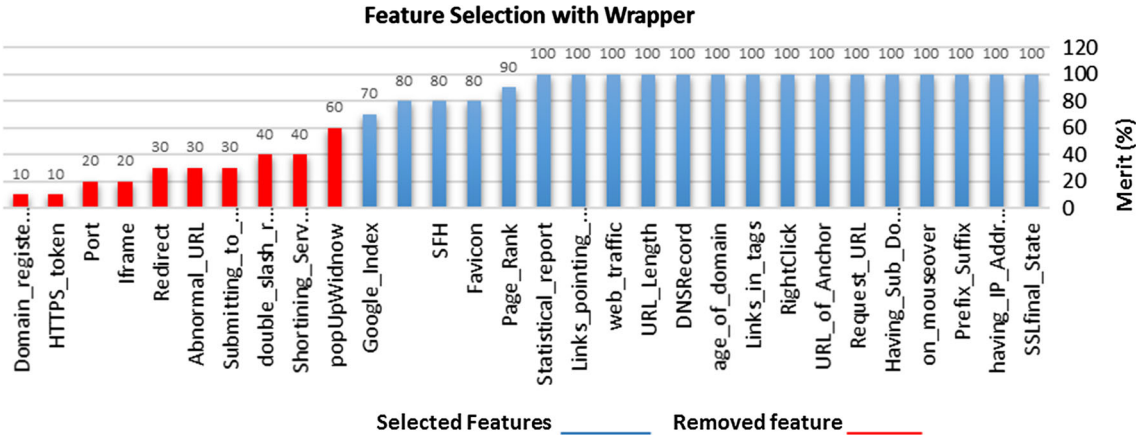


Fig. 7 Feature selection with wrapper

of each feature. The accuracy of feature selection methods is evaluated based on the values of precision and recall.

Precision is defined as the ratio of the number of correct phishing classes, toward of the phishing classes.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

where TP denotes the number of classes that are correctly labeled as phishing webpages and FP is the number of classes that are incorrectly labeled as phishing webpages.

Recall is defined as the ratio of the number of correct phishing classes to the sum of the corrected ones with the phishing websites which are misidentified as legitimated.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

where FN is the number of classes that are incorrectly labeled as legitimate webpage.

Table 3 Decision tree outputs

	Class = 1	Class = −1	Weighted avg.
TP Rate	0.971	0.941	0.958
FP Rate	0.059	0.021	0.045
Precision	0.954	0.963	0.958
Recall	0.971	0.941	0.958
F-measure	0.963	0.952	0.958

The F-measure is also defined as a measure of the test accuracy.

$$F\text{-measure} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Tables 3 and 4 list the obtained results for the DT and wrapper methods. It can be seen that the F-measure of the wrapper method is better than of the DT approach.

As listed in Tables 3 and 4, the accuracy of the wrapper method is equal to 96. 3% considering 20 selected

features. According to the results, by adding or removing one or more items from the features list, the accuracy is decreased. Hence, it can be concluded that feature selection using wrapper method has better performance in comparison with DT.

Figure 8 shows the accuracy of the DT evaluator for different feature numbers. As shown in this figure, the red point is the obtained highest accuracy of the DT, where the feature subset contains 20 members that this case gives a better result than the other subsets.

Once the best subset of the features is extracted, the NR based on the modified harmony trained the data to give the special target defined as the prediction of phishing websites with a high accuracy. The accuracy of the two methods, including NR and SVM, is then calculated using Eq. (13). The NR model outputs and a comparison to the SVM performance are illustrated in Table 4.

Table 4 Wrapper outputs

	Class = 1	Class = −1	Weighted Avg.
TP Rate	0.974	0.949	0.963
FP Rate	0.051	0.026	0.04
Precision	0.96	0.967	0.9632
Recall	0.974	0.949	0.963
F-measure	0.967	0.958	0.963

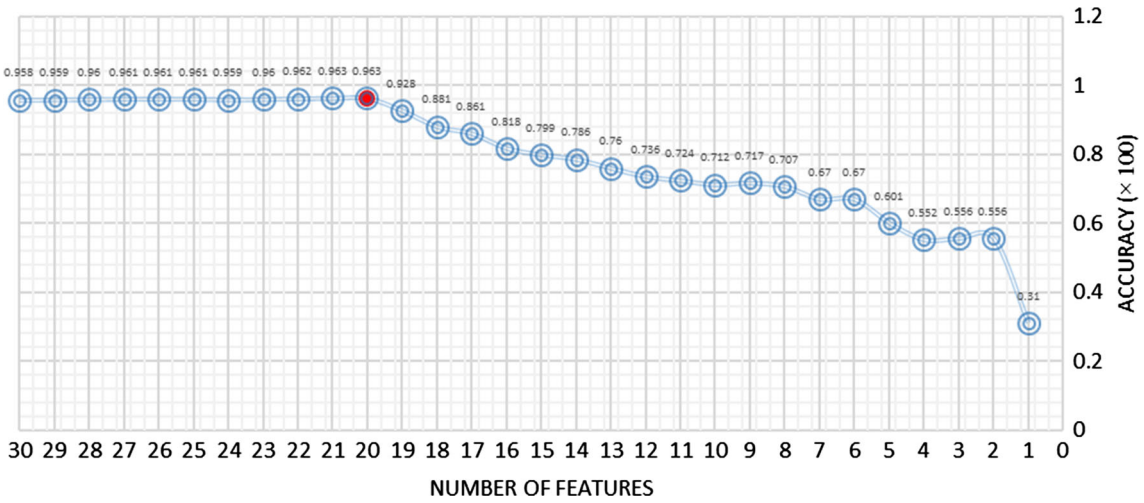


Fig. 8 Accuracy of DT in different number of features

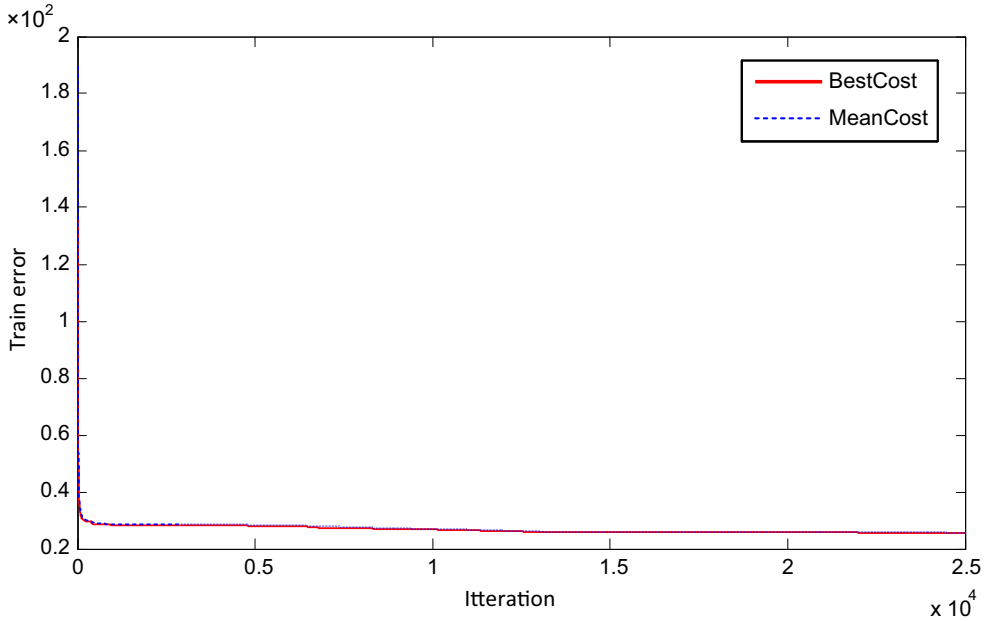


Fig. 9 Best-cost and mean costs of HS

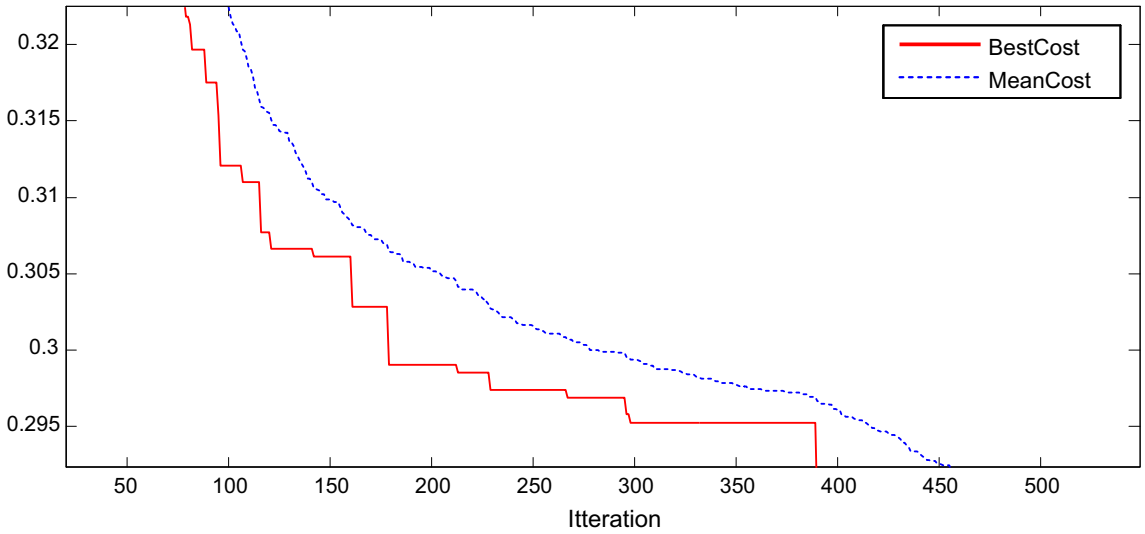


Fig. 10 Details of best-cost and mean-cost

Table 5 Accuracy of HS and SVM algorithms

	Train	Test
HS	94.1384	92.8087
SVM	92.578	91.8318

$$\text{Accuracy (train and test)} = \frac{N_{(\text{predict=real})}}{N_T} \times 100 \qquad (13)$$

where $N_{(\text{predict}=real)}$ and N_T are the number of instances that predicted class label which are equal to the desired class and total number of instance, respectively.

In the HS, the initial parameters can be listed as HMS = 30, maxItteration = 25000, $\text{PAR}_{\min} = 0.1$, $\text{PAR}_{\max} = 0.5$ and HMCR = 0.995. The PAR is changed dynamically in each iteration.

Figure 9 shows the best-cost and mean-cost of the harmony search. The best-cost is calculated by the best harmony and the mean-cost is evaluated using the average cost of all harmonies. According to this figure, the best-cost and mean-cost charts are not coincident. This fact is illustrated in Fig. 10 which is a large scale picture of Fig. 9.

The SVM algorithm is implemented in Matlab using the LIBSVM library (Li et al. 2016). LIBSVM is a free library which provides four basic kernels and implements a tool named "Cross-validation and Grid-search" to approximate the appropriate penalty parameters (C). In this library, the svmtrain function is used for training data and the svmpredict is used to predict the accuracy of testing and training data. The dataset is divided into three partition, and two-third of the dataset is used for training and one-third of it is used for testing. The members of the partitions are selected randomly. The results of Table 5 confirm that the NR-based HS algorithm introduces a better accuracy compared to the SVM in both train and test phases.

4.2 Comparative analysis

In order to verify the efficiency of the proposed method, some related researches are investigated for comparison. Mohammad et al. (2014b) have used 17 features which are evaluated with the self-structuring neural network. They have achieved 92.48% test accuracy and 93.45% train accuracy. Hamid and Abawajy (2011) examined 7 features generated using a hybrid-feature scheme as an indicator to specify the best classification method for phishing email detection. They have compared the accuracy of 4 classification methods in 3 datasets. The classification methods include bayes net (BN), decision tree, adaBoost and random forest (RF). As shown in their results, the adaBoost, RF and BN have a better performance in dataset 1, dataset 2 and dataset 3, respectively. Bottazzi et al. (2015) have presented a novel framework for phishing detection in mobile devices. The features used in their research are gathered from URL and HTML source of websites. They have used 4 classification methods for assessment of the accuracy of the framework. As shown in their results, the J48 method conducted a better performance in comparison with other algorithms. Here, the proposed detection method which is based on MHS and SVM is compared with the above-mentioned methods. The dataset which is used in our proposed method is similar to that of used in Mohammad et al. (2012). Table 6 confirms that the proposed method for phishing detection has high degree of efficiency than some of the previous mentioned methods.

Table 6 Comparison of different methods with our proposed method

Reference	Number of feature	Number of instances	Feature selection method	Accuracy
Mohammad et al. (2014b)	17	1400 (1120:280)	_	Self-structuring neural network (92.48%)
Hamid and Abawajy (2011)	7	6932	Hybrid-feature	Dataset1 (91%) Dataset2 (93%) Dataset3 (92%)
Bottazzi et al. (2015)	53	86000	_	J48 (89.2 %), Bayes net (78%) SMO (78.1%), SDG (79.6%)
Proposed method	20	11055 (7370:3685)	Wrapper	Harmony search (92.80%) SVM (91.83%)

5 Conclusion

In this paper, the procedure of phishing websites detection is investigated using feature selection methods and meta-heuristic algorithms. At first, the more efficient features are selected from the available dataset applying the feature selection methods. The mentioned dataset consists of 30 features which 20 of them are selected and used by two phishing detection methods. The detection methods are the modified harmony search based nonlinear regression and SVM. As shown in the results, the meta-heuristic algorithm confirms better accuracy in comparison with heuristic algorithms. Applying the meta-heuristic algorithm in phishing detection methods has not been analyzed yet. The main results of our study are listed below.

The main results of this research is fourfold as follows.

1. The decision tree (DT) is used as an evaluator for different number of features. As a result, the feature subset containing 20 members is better than the others.
2. The DT and wrapper methods are used to select the most important features. The wrapper method presents a better performance in comparison with the DT one. In both approaches, by adding or removing one or more items from the feature list, the accuracy is decreased.
3. Two algorithms (NR-HS and SVM) are employed to detect the phishing websites. It should be said that the nonlinear regression is used as a cost function for HS. This study establishes a comparison between NR-HS and SVM algorithms and as a result, the NR-HS has a greater amount of precision comparing SVM.
4. In the proposed NR-HS approach, the pitch adjustment rate (PAR) and generated new harmony (GNH) parameters are changed in each iteration and these variations prove a better accuracy in comparison to the traditional method's accuracy with constant PAR and GNH.

In addition, there are several lines of research arising from this work which should be pursued. Firstly, it will be interesting to consider parallel memory for HS to reduce the runtime. Secondly, the reliability analysis (Qiu et al. 2017a) of the proposed HS can be investigated in future work. The third interesting suggestion is working on HS parameters. Parameters (PAR and HMCR) of the proposed method can be calculated intelligently using heuristic and meta-heuristic algorithms.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Expert Syst Appl 41:5948–5959

Aburrous M, Hossain MA, Thabatah F, Dahal K (2008) Intelligent phishing website detection system using fuzzy techniques. In: 3rd international conference on information and communication technologies: from theory to applications. ICTTA 2008. IEEE, pp 1–6

Aburrous M, Hossain MA, Dahal K, Thabtah F (2010) Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Syst Appl 37:7913–7921

Ameli K, Alfi A, Aghaebrahimi M (2016) A fuzzy discrete harmony search algorithm applied to annual cost reduction in radial distribution systems. Eng Optim 48:1529–1549

Basnet R, Mukkamala S, Sung AH (2008) Detection of phishing attacks: a machine learning approach. In: Soft computing applications in industry. Springer, pp 373–383

Bottazzi G, Casalicchio E, Cingolani D, Marturana F, Piu M (2015) MP-Shield: a framework for phishing detection in mobile devices. In: 2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing (CIT/IUCC/DASC/PICOM). IEEE, pp 1977–1983

Cai C, Han L, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a

protein from its primary sequence. Nucleic Acids Res 31:3692–3697

Cao J, Li Q, Ji Y, He Y, Guo D (2016) Detection of forwarding-based malicious URLs in online social networks. Int J Parallel Prog 44:163–180

Fil BA, Korkmaz M, Özmetin C (2016) Application of nonlinear regression analysis for methyl violet (MV) dye adsorption from solutions onto illite clay. J Dispers Sci Technol 37:991–1001

Gupta R, Shukla PK (2015) System design, investigation and countermeasure of phishing attacks using data mining classification methods and its analysis. Int J Adv Sci Technol 78:29–40

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11:10–18

Hamid IRA, Abawajy J (2011) Phishing email feature selection approach. In: 2011 IEEE 10th international conference on trust, security and privacy in computing and communications. IEEE, pp 916–921

He Y-L, Wang X-Z, Huang JZ (2016) Fuzzy nonlinear regression analysis using a random weight network. Inf Sci 364:222–240

Jahn J (2017) Karush–Kuhn–Tucker conditions in set optimization. J Optim Theory Appl 172:707–725

Jeong SY, Koh YS, Dobbie G (2016) Phishing detection on twitter streams. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 141–153

Kalivarapu J, Jain S, Bag S (2016) An improved harmony search algorithm with dynamically varying bandwidth. Eng Optim 48:1091–1108

Lee KS, Geem ZW (2005) A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. Comput Methods Appl Mech Eng 194:3902–3933

Li K, Wang F, Zhang L (2016) A new algorithm for image recognition and classification based on improved Bag of Features algorithm. Opt Int J Light Electron Opt 127:4736–4740

Manjarres D, Landa-Torres I, Gil-Lopez S, Del Ser J, Bilbao MN, Salcedo-Sanz S, Geem ZW (2013) A survey on applications of the harmony search algorithm. Eng Appl Artif Intell 26:1818–1831

Mohammad RM, Thabtah F, McCluskey L (2012) An assessment of features related to phishing websites using an automated technique. In: 2012 international conference for internet technology and secured transactions. IEEE, pp 492–497

Mohammad RM, Thabtah F, McCluskey L (2014a) Intelligent rule-based phishing websites classification. IET Inf Secur 8:153–160

Mohammad RM, Thabtah F, McCluskey L (2014b) Predicting phishing websites based on self-structuring neural network. Neural Comput Appl 25:443–458

Mohammad R, Thabtah FA, McCluskey T (2015) Phishing websites Dataset

Montazer GA, ArabYarmohammadi S (2013) Identifying the critical indicators for phishing detection in Iranian e-banking system. In: 2013 5th conference on information and knowledge technology (IKT). IEEE, pp 107–112

Naik B, Nayak J, Behera HS, Abraham A (2016) A self adaptive harmony search based functional link higher order ANN for non-linear data classification. Neurocomputing 179:69–87

Pandey M, Ravi V (2012) Detecting phishing e-mails using text and data mining. In: 2012 IEEE international conference on computational intelligence & computing research (ICCIC). IEEE, pp 1–6

Qiu J, Wei Y, Karimi HR, Gao H (2017a) Reliable control of discrete-time piecewise-affine time-delay systems via output feedback. IEEE Trans Reliab 99:1–13

Qiu J, Wei Y, Wu L (2017b) A novel approach to reliable control of piecewise affine systems with actuator faults. IEEE Trans Circuits Syst II Express Briefs 64:957–961

Rodrigues D, Pereira LA, Nakamura RY, Costa KA, Yang X-S, Souza AN, Papa JP (2014) A wrapper approach for feature selection based on bat algorithm and optimum-path forest. Expert Syst Appl 41:2250–2258

Satapathy SC, Chittineni S, Krishna SM, Murthy J, Reddy PP (2012) Kalman particle swarm optimized polynomials for data classification. Appl Math Model 36:115–126

Song Q, Jiang H, Liu J (2017) Feature selection based on FDA and F-score for multi-class classification. Expert Syst Appl 81:22–27

Wang L, Ni H, Yang R, Pappu V, Fenn MB, Pardalos PM (2014) Feature selection based on meta-heuristics for biomedicine. Optim Methods Softw 29:703–719

Wang G-G, Gandomi AH, Zhao X, Chu HCE (2016) Hybridizing harmony search algorithm with cuckoo search for global numerical optimization. Soft Comput 20:273–285

Wei Y, Qiu J, Karimi HR (2017) Reliable output feedback control of discrete-time fuzzy affine systems with actuator faults. IEEE Trans Circuits Syst I Regul Pap 64:170–181

Xia Z, Wang X, Sun X, Liu Q, Xiong N (2016) Steganalysis of LSB matching using differences between nonadjacent pixels. Multimed Tools Appl 75:1947–1962