International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5380-5390

https://doi.org/10.70517/ijhsa464377

Research on Improved Decision Tree Algorithms for Identifying Phishing Attacks

Haoran Yang[1,*], Yi Li[2], Chang Liu[3] and Yichuan Zhou[4]

[1] Beijing Troy Cloud Data Technology Co., Ltd., Beijing, 100071, China
[2] School of Computer Science and Technology, Jilin University, Beijing, 100010, China
[3] Department of Hospitality and Business Management, The Technological and Higher Education Institute of Hong Kong, Hong Kong, 999077, China
[4] Shanghai Shiyun Information Technology Co., Ltd., Shanghai, 200120, China

Corresponding authors: (e-mail: admin@adysec.com).

Abstract Under the rapid progress of Internet technology, phishing attacks have become a serious threat in network security. However, traditional decision tree algorithms often encounter the dual difficulties of unstable classification accuracy and low computational efficiency in recognizing such attacks. To address this problem, this paper focuses on creating an improved C4.5 decision tree algorithm that integrates the boundary point principle with learning vector quantization. By virtue of the boundary point principle, this algorithm effectively reduces the number of candidate segmentation points and greatly improves the efficiency of the algorithm. On top of that, the learning vector quantization approach is introduced to intelligently cluster the raw data, which in turn optimizes the segmentation point selection mechanism. More importantly, by deeply integrating the information entropy and covariance characteristics together, a set of attribute selection mechanism with higher accuracy is constructed. The results show that the proposed algorithm exhibits excellent classification performance when dealing with a wide range of datasets. Especially when dealing with high-dimensional and complex data environments, not only the classification accuracy is significantly improved, but also the computational efficiency has an essential leap. This study provides an efficient and accurate solution for phishing attack identification, which not only has the value of theoretical research, but also presents a broad application prospect and far-reaching social significance in practical application.

Index Terms phishing attack, decision tree algorithm, boundary point principle, learning vector quantization, information entropy, covariance

I. Introduction

I. A. Background and significance of the study

Accompanied by the rapid development and widespread popularization of Internet technology, the network has been deeply integrated into all aspects of people's daily life and work. Nowadays, the network environment presents a complex and varied situation, network security-related issues have become more and more prominent, and phishing attacks, a highly deceptive way, have evolved into a serious threat that cannot be taken lightly in the context of network security. Such attacks often disguise themselves as trustworthy entities, leading users to hand over sensitive information such as usernames, passwords, and bank account numbers, or luring users to click on links and attachments with malicious programs, thus achieving the purpose of stealing users' private data and even controlling their devices. The China Internet Network Information Center's “Statistical Report on China's Internet Development” clearly shows that phishing attacks have risen to become one of the major threats to users' network security, with the number of attacks rising year after year and the means continuously being renovated. The rapid rise of e-commerce, coupled with the widespread use of online payment, has pushed the phishing attacks to show diversified, precise and covert characteristics. Attackers not only build phishing websites that are very similar to regular websites, but also use social engineering to carry out targeted attacks, and even automatically generate unrecognizable phishing content with the help of artificial intelligence technology. The traditional identification method of relying on blacklists and feature matching is not sufficient in the face of these new attacks, and there is an urgent need to develop smarter and more efficient identification techniques. Among the many machine learning algorithms, decision trees are widely used in the field of phishing attack identification due to their strong interpretability, high computational efficiency and outstanding classification ability. However, traditional decision tree algorithms, such as ID3 and C4.5, still have many shortcomings when dealing with high-dimensional large-scale data, including unstable classification accuracy, inefficient algorithm execution, and easy to overfitting and other problems.

As a classic classification algorithm, the core of the decision tree algorithm is to build a tree structure model to establish a mapping relationship between data features and decision results. In the process of phishing attack identification, the decision tree can effectively extract URL features, web page content features and network behavior features and other related information for classification judgment. However, in the face of the complexity and diversity of phishing attacks, the traditional decision tree algorithm has obvious defects in the discretization of continuous attributes, and its efficiency is relatively low when dealing with high-dimensional data. Therefore, the research on phishing attack identification technology has important theoretical and practical significance. From the theoretical level, the improvement of the decision tree algorithm enriches the application theory of machine learning in the field of network security, and provides new ideas for subsequent research. From a practical point of view, efficient and accurate phishing attack identification technology can effectively protect the safety of user information, reduce economic losses, and is of great significance for building a safe and reliable network environment. With the continuous evolution of phishing attack technology, the research on phishing attack identification mechanism based on improved decision tree algorithm has far-reaching practical significance for improving the ability of network security protection.

I. B. Main contributions and innovations of this study

This study focuses on solving the efficiency and accuracy problems of traditional decision tree algorithms for phishing attack identification, and proposes an optimization scheme with multidimensional characteristics. In the process of continuous attribute discretization, we make innovative improvements to the C4.5 decision tree algorithm by introducing the boundary point principle to filter the candidate segmentation points. The traditional C4.5 algorithm needs to traverse all possible segmentation points and calculate the information gain rate, which is a heavy computational burden in the context of high-dimensional big data. Based on Fayyad's Boundary Point Determination Theorem, this study proposes that only when neighboring samples belong to different categories, the segmentation points between them have the possibility of becoming the optimal choice. By identifying and retaining these critical boundary points as candidates, and eliminating the useless segmentation points, the execution time of the algorithm can be shortened with only a small loss of classification accuracy, which demonstrates a very good ratio of time-to-space efficiency. On this basis, the learning vector quantization method is innovatively combined with the boundary point principle to construct a more accurate segmentation point selection mechanism. The traditional decision tree often ignores the internal clustering structure of the data, which leads to a less than ideal selection of segmentation points. In this study, the LVQ algorithm is used to cluster the data, find the approximate boundaries of different categories of data, and add them as candidate segmentation points, which makes the classification decision more reasonable, fully utilizes the information contained in the internal structure of the data, and effectively improves the accuracy of classification while ensuring the efficient operation of the algorithm.

Another core innovation of this study is the organic integration of information entropy and covariance to construct a more accurate attribute selection mechanism. The traditional decision tree mainly relies on the information gain or gain rate to determine the split attributes, which fails to fully consider the interrelated characteristics of attributes. To address this shortcoming, this paper proposes a coordinated measure of information entropy method, which not only considers the amount of information carried by individual attributes, but also incorporates the covariance relationship between attributes, so that attribute features with more discriminatory ability can be selected in the process of constructing decision trees. This improvement makes the generated decision tree structure more reasonable, effectively reduces the risk of overfitting, and enhances the generalization ability of the model in identifying various kinds of complex phishing attacks. Experimentally, this method has achieved better recognition results than the traditional algorithms on several phishing datasets, especially when dealing with complex samples that contain multiple attack features. These three technical improvements - boundary point filtering mechanism, learning vector quantitative clustering optimization, and information entropy and covariance combination - comprehensively improve the performance of the decision tree algorithm in phishing attack identification. These improvements not only enrich the decision tree algorithm at the theoretical level, but also provide practical technical solutions for the construction of an efficient and accurate phishing attack identification system in practical applications.

II. Literature review

II. A. Theoretical basis of the study

Decision trees, as a class of classification models with a tree-like architecture, are used to build predictive models by iteratively classifying the characteristics of samples. In this process, it consists of three basic structures: root nodes, internal nodes, and leaf nodes. The root node symbolizes the overall set of samples, the internal nodes

represent the testing of a particular characteristic, and the leaf nodes represent the final classification results. The process of building a decision tree is essentially to continue to select the most suitable characteristics to be split, until it meets the stopping conditions. Information entropy plays a key role in the decision tree algorithm to measure the purity of data, which is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}$$

where $p(x_i)$ represents the probability of category i in the sample set.

A smaller value of information entropy of the data set represents higher purity and vice versa indicates higher uncertainty.The ID3 algorithm is an early decision tree algorithm that uses information gain as an attribute selection criterion, i.e.:

$$Gain(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{2}$$

where S represents the set of samples, A represents the attribute to be tested, $Values(A)$ represents all possible values of attribute A, and S_v represents the subset of samples in which attribute A takes the value of v. The ID3 algorithm selects the attribute with the largest information gain as the split attribute each time to construct the decision tree, but it has the defects of being biased toward multi-valued attributes and unable to deal with the continuous type attributes and missing values directly. Algorithm C4.5 improves ID3 by introducing the concept of information gain rate to solve the problem of biased multi-valued attributes, which is defined as:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \tag{3}$$

$$SplitInfo(S, A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|} \tag{4}$$

$SplitInfo(S, A)$ denotes the entropy of attribute A. C4.5 can also process continuous attributes and discretize continuous attributes by finding the best split point. When dealing with continuous attributes, C4.5 will sort the samples by attribute values, consider the midpoints between neighboring samples as possible segmentation points, and select the point with the largest rate of information gain, but this method has high computational complexity and is inefficient in dealing with high-dimensional big data.

The CART algorithm uses the Gini index as the attribute selection criterion for classification and regression tasks, and the Gini index is defined as:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2 \tag{5}$$

where p_i denotes the proportion of category i in the sample set.CART selects the attribute that minimizes the Gini index after division each time for splitting to construct a binary decision tree.

The application of covariance in decision tree algorithms has become a hot research topic in recent years. which measures the degree of linear correlation between two random variables. that is:

$$Cov(X, Y) = E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right] \tag{6}$$

The introduction of covariance values can better capture the interrelated status of attributes, and then select attributes that are more capable of differentiation. Chen Liang proposed a decision tree algorithm based on information entropy and covariance, which improves the accuracy of classification by constructing logical connections between multidimensional data.

The boundary point principle belongs to a theory of optimizing the discretization of continuous attributes, which suggests that in binary classification problems, the optimal segmentation points must be located at the boundary of different categories of samples. Related research applies the boundary point principle to the C4.5 algorithm, which significantly reduces the number of candidate segmentation points and improves the efficiency of the algorithm. Learning vector quantization is a neural network model based on competitive learning, which is able to

perform effective clustering operations on data. Combining the learning vector quantization with the boundary point principle can more accurately identify the boundary situation between categories and optimize the selection process of segmentation points. The information dispersion index is used to penalize multi-valued attributes, which can reflect the degree of multi-value of the attributes and reduce the dependence of the evaluation criteria on multi-valued features. The traditional decision tree algorithm has deficiencies in dealing with high-dimensional big data and discretization of continuous attributes. By introducing the principle of boundary point, quantization of learning vectors, and the combination of information entropy and covariance, the decision tree algorithm is able to effectively improve the accuracy ratio and the efficiency status of the decision tree algorithm in phishing attack identification. These theories provide a solid supporting foundation for the improved algorithm proposed in this paper.

II. B. Status of research

Phishing is a type of cyber attack that obtains sensitive information of users through false pretenses, and it is a major threat in the current cyber security field [1]. Phishing attacks utilize social engineering techniques by sending fake emails, text messages or links to lure users to click on them in order to obtain sensitive information such as account passwords, banking information, ID numbers, etc [2]-[4]. In response to such attacks, it is particularly important to study the identification mechanism of phishing attacks [5].

Currently, phishing attacks are becoming increasingly sophisticated, and the attackers' attack techniques are constantly updated [6]. Traditional defense means such as anti-spam system, malicious link detection, etc. have been difficult to cope with the new phishing attacks, and the phishing attack identification mechanism based on decision tree algorithm is a key technology emerging in this context [7]-[9]. Decision tree is a classification method commonly used in machine learning and data mining, and its model is similar to a tree with nodes representing attribute tests [10].

At each node, we select a branch based on the value of an attribute until we reach the leaf node, which represents the classification result [11], [12]. The process of building a decision tree is to classify the data by dividing the attributes so that the samples corresponding to each branch node have the same attribute values [13]-[15]. In phishing attack recognition, decision tree recognition is one of the important topics in the current cyber security field [16], [17]. This technique uses the characteristics of network traffic data to achieve the detection and defense of network attack behaviors by constructing a decision tree model, although it faces some challenges, the classification performance and accuracy of the decision tree model can be further improved through the application of improved algorithms and techniques [18]-[21].

III. Phishing attack identification mechanism based on improved decision tree algorithm

III. A. Improved algorithm design

This study focuses on the efficiency and accuracy problems of the traditional C4.5 decision tree algorithm for phishing attack identification, and provides a set of comprehensive improvement initiatives in an innovative way. The solution cleverly integrates the boundary point principle, learning vector quantization, and information entropy and covariance properties to construct a multi-level optimization architecture, which greatly improves the effectiveness of decision trees in phishing identification.

The C4.5 algorithm, when dealing with high-dimensional continuous attribute data such as phishing URL features, must calculate the information gain rate for all possible segmentation points, and its computational overhead is extremely huge, which restricts the popularization of the algorithm in practical applications. To address this problem, this paper screens the candidate segmentation points for optimization based on Fayyad's boundary point determination theorem, which shows that the optimal segmentation points must be at the boundaries of different categories of samples in the binary classification problem, formally expressed as follows:

If the values of continuous attribute A have been arranged in ascending order $\{a_1, a_2, \ldots, a_n\}$, and the corresponding category label is $\{c_1, c_2, \ldots, c_n\}$, then:

$$BP = \{(a_i + a_{i+1})/2 \mid c_i \neq c_{i+1}, 1 \leq i < n\} \tag{7}$$

where BP is the set of candidate segmentation points; by considering only the midpoints at the category changes as candidate points, the computational amount is greatly reduced, which is particularly effective in phishing attack feature recognition. Simply using the boundary point principle may ignore the internal structure of the data information leading to a slight decline in classification accuracy, in order to compensate for this shortcoming, the introduction of learning vector quantization method for data clustering, to further optimize the selection of

segmentation points.LVQ algorithm through the competitive learning mechanism to find out the prototype vectors of each class, which represent the internal distribution characteristics of the data, the core steps include:

Initialization for each class randomly selects samples as prototype vectors, competitive learning to find the closest prototype vector w_c to the training sample x, update the prototype vector, i.e.:

$$w_c(t+1) = \{w_c(t) + \alpha(t)[x - w_c(t)]\} \tag{8}$$

where x and w_c belong to the same class, and if x and w_c belong to different classes, $\alpha(t)$ is the learning rate that decreases with time t. Repeat the above steps until convergence. After obtaining the prototype vectors, calculate the midpoint between the prototype vectors of different classes as the candidate splitting point, i.e:

$$BP_{LVQ} = \{(w_i + w_j)/2 \mid class(w_i) \neq class(w_j), \forall i, j\} \tag{9}$$

The final set of candidate segmentation points is the concatenation of the set of boundary points and the set of LVQ-generated segmentation points, i.e:

$$BP_{final} = BP \cup BP_{LVQ} \tag{10}$$

This approach fully utilizes the intrinsic structural information of the data to make the selection of split points more reasonable, which is especially suitable for dealing with feature-type data with complicated distribution in phishing attacks.

The traditional C4.5 algorithm selects the split attributes only by the information gain rate, but does not consider the correlation between the attributes. However, there are often complex correlations between the features of phishing attacks, e.g., the length of the domain name and the number of special characters in the URL are correlated with each other. In order to solve this problem, this study introduces the covariance property to construct a coordinated measure of information entropy to optimize the attribute selection process, which is defined as:

$$CME(A) = Gain\_Ratio(A) \times (1 + \lambda \times Corr\_Factor(A)) \tag{11}$$

where $Gain\_Ratio(A)$ is the information gain rate of attribute A, $Corr\_Factor(A)$ is the correlation factor between attribute A and other attributes, λ is the weight parameter, and the correlation factor is computed as:

$$Corr\_Factor(A) = \sum_{B \neq A} |Corr(A,B)| \times Gain\_Ratio(B) \tag{12}$$

where $Corr(A,B)$ is the attribute correlation coefficient: $Corr(A,B) = \dfrac{Cov(A,B)}{\sigma_A \sigma_B}$, $Cov(A,B)$ is the covariance, and σ_A and σ_B are standard deviations, respectively.

The information entropy of the measurement coordination not only pays attention to the distinguishing ability of the attribute itself, but also considers the degree of its association with other attributes, so as to enable the decision tree to select a combination of attributes with more distinguishing ability. Based on the above three optimization operations, the improved C4.5 decision tree algorithm process covers: phishing attack feature data are standardized in the data preprocessing stage; for each consecutive attribute, the boundary point principle is applied to screen out the initial candidate segmentation points. The LVQ algorithm is used to cluster the data to generate the category prototype vectors, calculate the midpoints between the prototype vectors of different categories and add them as candidate segmentation points, and calculate the information gain rate for each candidate segmentation point to select the optimal segmentation point. Calculate the information entropy of each attribute that makes the measure coordinated, select the optimal splitting attribute, split the data set based on the selected attributes and splitting points, and for the subset after splitting, execute the above steps repeatedly until the stopping condition is satisfied, and generate the decision tree model for phishing attack identification. The advantages of the improved algorithm in phishing attack recognition are as follows: the boundary point principle greatly reduces the number of candidate splitting points, which improves the efficiency of the algorithm; the optimization of the LVQ clustering makes the selection of splitting points more reasonable, which improves the classification accuracy; the information entropy that makes the measurement coordinated makes the selection of attributes more accurate, which strengthens the recognition ability of the model.

III. B. Experimental design and data set

In order to confirm the effectiveness of the optimized decision tree algorithm in phishing attack identification, we constructed a set of experimental planning, which covers five key aspects: data set selection, data pre-processing, experimental environment setting, evaluation index and comparison experiment design. In view of the wide applicability of the decision tree algorithm to various classification problems, five representative datasets from the UCI machine learning library and the KDD Cup 1999 dataset, which is a classic in the field of cybersecurity, are selected to evaluate the performance of the algorithm in the experimental process.

The UCI dataset is mainly used to verify the generalized performance of the improved algorithm in different classification tasks, while the KDD Cup 1999 dataset is tested in the cybersecurity environment to evaluate the effectiveness of the algorithm in recognizing phishing attacks. Table 1 presents the basic information of the UCI dataset used in the experimental process, which covers the data size, the number of attributes, the number of categories, and the data distribution characteristics.

Table 1: The UCI dataset has been traced

Dataset name	Sample size	Number of attributes	Number of categories	Data distribution characteristics
Iris	150	4	3	Balanced distribution, with 50 samples in each category
Wine	178	13	3	The distribution is unbalanced, with a category ratio of approximately 33:40:27
Banknote Authentication	1372	4	2	The distribution is close to equilibrium, with genuine banknotes: counterfeit banknotes 55:45
SPECTF Heart	267	44	2	Unbalanced distribution, normal: abnormal 60:40
MAGIC Gamma Telescope	19020	10	2	Unbalanced distribution, gamma rays: Background radiation 35:65

The KDD Cup 1999 dataset, as a standardized test collection in the field of network security, records a large number of network intrusions and covers many phishing attack characteristics, which is very suitable for verifying the phishing attack identification system. The data set has about 4.9 million records, each record consists of 41 characteristic attributes and 1 labeled attribute, Table 2 is the description of the KDD Cup 1999 data set.

Table 2: KDD Cup 1999 dataset

Attribute	Detailed information
Dataset scale	Approximately 4.9 million records (complete set), approximately 490,000 records (10% subset)
Number of features	41 (including 9 category features and 32 consecutive features)
Attack type	There are 22 specific attack types including normal traffic and four major types of attacks (DoS, Probe, R2L, U2R)
Related to phishing	Mainly in the R2L (remote to local) attack category, it includes features such as phishing emails and deceptive logins
Sample distribution	Extremely unbalanced, normal: Attacks 20:80, phishing related attacks account for approximately 0.1%
Characteristic attribute	It includes network characteristics such as duration, protocol type, source/destination byte count, and connection status

Given the large size of the initial KDD Cup dataset, we selected a 10-percent sampling subset for testing and ensured that the distribution of sample categories was consistent with the initial dataset. We systematically processed the initial data, including feature selection, noise handling, missing value response, and data normalization. For the UCI dataset, all initial features are retained. For the KDD Cup dataset, 12 features closely related to phishing attacks are selected based on professional knowledge and statistical analysis. Noise disposal uses the Z-score method to identify abnormal values, marking sample points with an absolute Z-score greater than 3 as potential noise points, eliminating data that is obviously erroneous, and retaining abnormal values that may be valid. Missing values should be dealt with according to the proportion of missing to take different strategies:

When the percentage of missing values is less than five percent, numerical features are filled with means and categorical features are filled with plurals. When the missing percentage is between five and thirty percent, the KNN filling algorithm based on attribute correlation is used. When the missing percentage of a single feature is more than thirty percent, the feature is considered for deletion. Continuous features are normalized by min-max normalization to ensure that all feature values are mapped to intervals to avoid decision bias due to differences in magnitude.

The experimental environment used Intel Core i7 - 10700K processor, 16GB RAM, Python 3.8 programming environment and scikit - learn 0.24.2 machine learning library as the infrastructure. We evaluate the algorithm performance in all aspects through accuracy, precision, recall, F1 score, and ROC - AUC values, and record the execution time to assess the computational efficiency. Ten-fold cross-validation is used to guarantee the reliability of the results by randomly dividing the data into 10 parts, sequentially using 9 parts as training and 1 part as testing, and taking the average performance metrics as the assessment of the algorithm performance. The experiments compare the traditional C4.5 algorithm, the improved C4.5 algorithm based on the boundary point principle (C4.5 - BP), the improved C4.5 algorithm based on learning vector quantization (C4.5 - LVQ), and the comprehensive improved C4.5 algorithm (C4.5 - BP - LVQ - CME) proposed in this paper, and verify the effects of the improvement points on the algorithm's performance step by step by using the control variable method. To guarantee the fairness of the experiment, all comparison algorithms are run under the same computing environment, with the same data preprocessing method as well as cross-validation splitting. The phishing attack identification experiments for the KDD Cup dataset use a binary classification strategy, where phishing-related attacks are labeled as positive classes and the rest are labeled as negative classes. Given the extreme data imbalance, we apply stratified sampling as well as SMOTE oversampling techniques to balance the class distribution of the training set so that the model pays more attention to the minority class samples. In terms of parameter setting, the number of prototype vectors of the learning vector quantization algorithm is set to 5 times the number of categories, the initial value of the learning rate is set to 0.1, and an exponential decay strategy is used. The weight parameter in the coordinated measure information entropy λ determines the optimal value through a grid search, with a search range between 0.1 and 1.0. We use tables and charts to visualize the experimental results, clearly reflecting the performance differences of the algorithms on different datasets and providing data support for analyzing the advantages of the algorithms.

III. C. Experimental results and analysis

In this section, a series of controlled experiments are conducted to evaluate the performance of the improved algorithm in phishing attack identification. The experimental results clearly show the advantages of the modified C4.5 decision tree algorithm based on the boundary point principle and learning vector quantization in terms of classification accuracy and computational efficiency. The control results of classification accuracy are given in Figure 1.

It can be clearly observed that, as the complexity of the dataset increases, the advantages of the comprehensive and improved algorithm proposed in this paper become more and more significant. Especially on the two high-dimensional datasets, SPECTF Heart and MAGIC Gamma Telescope, the accuracy of the improved algorithm increases to the greatest extent, reaching 4.12% and 3.64%, respectively. This indicates that the algorithm has stronger robustness and generalization ability in dealing with high-dimensional features, which is extremely suitable for high-dimensional feature classification tasks such as phishing attack identification.

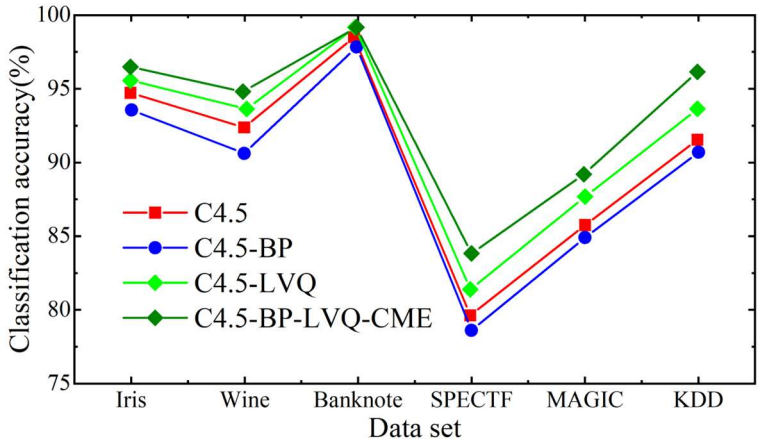

Figure 1: Comparison of classification accuracy of different algorithms

The efficiency of the algorithms is another key element in measuring the performance of decision tree algorithms, and Table 3 presents a comparison of the runtime of each algorithm on different data sets. As can be seen from the data in the table, the C4.5 - BP algorithm based on the boundary point criterion significantly reduces the runtime on all data sets, by an average of about 53%. This fully confirms the significant utility of the

boundary point criterion in reducing the number of candidate segmentation points and improving the efficiency of the algorithm. The C4.5 - LVQ algorithm, on the other hand, has a slightly higher runtime than the traditional C4.5 algorithm due to the introduction of the Learning Vector Quantization clustering process. Despite the integration of the three improved techniques, the integrated and improved algorithm C4.5 - BP - LVQ - CME proposed in this paper still reduces the overall runtime by about 45% compared with the traditional C4.5 algorithm through the efficient filtering of the boundary point criterion, showing a good degree of computational efficiency. Especially when dealing with large-scale data such as the KDD Cup phishing dataset, the runtime is reduced from 28.436 seconds to 15.682 seconds, which is a significant improvement in efficiency.

Table 3: The execution time of different algorithms is compared (s)

Dataset	C4.5	C4.5-BP	C4.5-LVQ	C4.5-BP-LVQ-CME
Iris	0.025	0.014	0.032	0.021
Wine	0.048	0.026	0.057	0.035
Banknote Authentication	0.187	0.092	0.215	0.124
SPECTF Heart	0.763	0.362	0.825	0.415
MAGIC Gamma Telescope	5.842	2.731	6.324	3.175
KDD Cup	28.436	13.257	30.845	15.682

Figure 2 shows a comparison of the execution time of various algorithms on data sets of different sizes. From this figure, it can be clearly perceived that along with the increasing size of the data set, the efficiency improvement brought by the boundary point principle becomes more and more prominent and significant. In smaller data sets such as the Iris data set, the difference in the execution time of various algorithms is not large; however, in larger data sets such as the KDD Cup data set, the improved algorithms have an extremely obvious advantage in terms of efficiency. This means that the improved algorithm proposed in this paper is especially suitable for processing large-scale phishing attack data information, and can fully meet the requirements of real-time in the process of practical application. Further analyzing the experimental results of phishing attack identification, we notice that the improved algorithm also achieves significant improvement in the precision rate, recall rate and F1 score. On the KDD Cup dataset, the C4.5 - BP - LVQ - CME algorithm achieves a precision rate of 94.83%, a recall rate of 96.57%, and an F1 score of 95.69%, which are all higher than the other algorithms used for comparison. This shows that the improved algorithm is not only able to accurately identify phishing attacks, but also able to effectively reduce the false alarm rate, which has high practical application value. It should be noted that the advantages of the improved algorithm are more significant when dealing with high dimensional data. In the case of SPECTF Heart (44 dimensions) and KDD Cup phishing data (41 dimensions), the accuracy is improved by 4.12% and 4.48%, respectively. This is mainly due to the effective use of the coordinated measure of information entropy on the correlation between attributes, which enables the algorithm to select more distinguishable combinations of attributes, thus improving the accuracy of classification. At the same time, the organic combination of the boundary point principle and learning vector quantization makes the selection of segmentation points more reasonable and appropriate, which further enhances the performance of the algorithm on high dimensional data. Comprehensive experimental results show that the improved C4.5 decision tree algorithm based on the boundary point principle and learning vector quantization proposed in this paper has significant advantages in the field of phishing attack identification, which not only improves the accuracy of classification, but also greatly reduces the computational complexity, and provides strong support for the construction of an efficient and accurate phishing attack identification system.

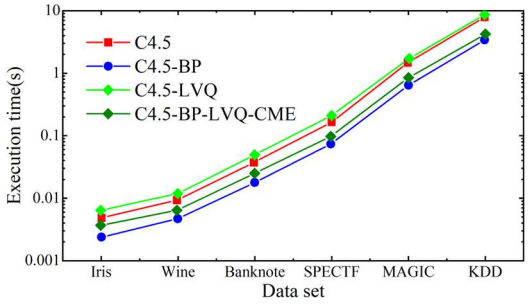

Figure 2: The execution time of different algorithms is compared (logarithmic scale)

IV. Conclusions and outlook

IV. A. Conclusions of the study

In this paper, a modified C4.5 decision tree algorithm based on the principle of boundary point and learning vector quantization is presented to address the lack of classification accuracy and poor computational efficiency of traditional decision tree algorithms in phishing attack identification. After systematic theoretical analysis and multiple rounds of experiments, this study finds that:

Based on the boundary point principle, the candidate segmentation point screening strategy makes the algorithm more efficient, and this optimization scheme reduces the execution time of the algorithm by 53% on average, and the execution time on the KDD Cup data set is drastically reduced from 28.436 seconds to 13.257 seconds, while the classification accuracy is generally stable. This improvement enables the algorithm to be more agile in responding to phishing threats, which meets the stringent real-time requirements of real-world applications. The segmentation point optimization strategy based on learning vector quantization improves the classification accuracy of the algorithm by calculating the midpoints between different categories of prototype vectors as the candidate segmentation points, making full use of the internal structure information of the data. The introduction of coordinated measure of information entropy further strengthens the algorithm's ability to control the complex association between phishing attack features, enabling the decision tree to select more distinguishable combinations of attributes.

It should be noted that the improved algorithm shows more prominent advantages in dealing with high-dimensional complex data, and the accuracy improvement is as high as 4.12% and 4.48% in the two high-dimensional data sets of SPECTF Heart (44 dimensions) and KDD Cup phishing data (41 dimensions), respectively. This fully indicates that the improved algorithm proposed in this paper is particularly suitable for phishing attacks, which is a high-dimensional feature classification task with strong robustness and generalization performance. Although the improved algorithm C4.5 - BP - LVQ - CME combines three improved techniques, the overall execution time is still reduced by about 45% compared with the traditional C4.5 algorithm by virtue of the efficient screening mechanism of the boundary point principle. At the same time, the classification accuracy is significantly improved, which strikes a good balance between efficiency and accuracy, and makes the improved algorithm show high application value in the actual network security defense system. The improved C4.5 decision tree algorithm based on the boundary point principle and learning vector quantization proposed in this study has significant advantages in the field of phishing attack identification, which can effectively identify various types of phishing attacks by improving the classification accuracy and significantly reducing the computational complexity. Enhancing the defense capability of network security system is of great significance for building a safer and more reliable network environment, and the improved ideas and methods proposed in this study also provide valuable reference for other machine learning algorithms in the field of network security, which has high theoretical value and practical value.

IV. B. Research limitations and future prospects

This academic research, although it has achieved a certain degree of success, still encounters a number of limitations and challenges that should not be taken lightly. In terms of data diversification, although we have conducted extensive tests on the UCI Machine Learning Library and the KDD Cup 1999 data set, these well-constructed lab data cannot completely reflect the complex and variable nature of real phishing attacks. In the real network environment, phishing methods are constantly updated, such as social engineering phishing, targeted phishing and other advanced categories, often with strong stealth characteristics and contextual correlation characteristics, only by relying on static feature analysis, it is difficult to effectively identify these dynamically evolving attack modes, especially for the identification of the zero-day attack is still weak. The stability of the algorithm under specific data characteristics is also problematic. When faced with extremely uneven distribution of data sets (e.g., phishing samples occupy only 0.1% of the KDD Cup data), even if the SMOTE oversampling technique is used for balancing, the artificially constructed samples may introduce additional interference, which may result in weakening of the model's generalization ability in the real environment. The generalization ability of the model in the real environment is weakened. Experimentally verified, the algorithm's performance decreases significantly when dealing with large-scale and high-dimensional data, and it is observed that when the number of samples exceeds 1 million and the feature dimension is greater than 50, the algorithm's execution time still exhibits a super-linear growth even when boundary point optimization measures are used, which poses a challenge for the real-time detection system.

Regarding the sensitivity of the weight parameter in the coordinated measurement of information entropy, we find that the optimal values of different data sets vary greatly (between 0.2 and 0.8), and the lack of adaptive adjustment mechanism greatly limits the flexibility of the algorithm in a variety of scenarios. When facing new

phishing attacks, the algorithm needs to retrain the whole model, which is inefficient and difficult to meet the demand for real-time defense. Although the decision tree itself has good interpretable characteristics, after we comprehensively improve the algorithm, the decision basis becomes more complex, and the lack of intuitive visualization tools to assist security analysts to understand the model decision process, which to a certain extent affects the degree of credibility of the system as well as its practical value. Together, these problems constitute a major obstacle to the current research and need to be overcome in future work. The research results also show that the algorithms are not flexible enough to capture the evolution of attack patterns when dealing with dynamically changing attack characteristics, which has a considerable limitation on the effectiveness of defense in the practical application environment.

The trend of future research can be promoted and improved from multiple dimensions, and the construction of a more comprehensive and real phishing attack data collection should be in the first place, which needs to cover the latest attack methods and technical characteristics, so as to provide a more realistic test environment for algorithm evaluation. Combining improved algorithms with deep learning techniques can be a convenient way to extract the advanced semantic features of phishing URLs with the help of convolutional neural networks or recurrent neural networks, which can make up for the deficiencies of traditional decision tree algorithms in dealing with sequential data and implicit features. The introduction of advanced integrated learning architectures such as Random Forest or XGBoost, which integrate the advantages of multiple improved decision trees, can undoubtedly further improve the classification performance and robustness. The problem of computational complexity also needs to be addressed by parallelization and distributed computing strategies, especially GPU-based parallelization, which is promising to achieve orders of magnitude performance improvement when dealing with very large data. The research and development of adaptive parameter adjustment strategy is also very critical, based on the data characteristics of the automatic optimization of weight parameters, can effectively reduce the burden of manual adjustment of parameters. Explore the lightweight implementation of algorithms in mobile devices and the Internet of Things environment, to address the security challenges of resource-constrained devices, and combined with emerging technologies such as blockchain to build a distributed security defense system based on improved decision trees, to achieve multi-node collaborative detection and defense, which will create a more powerful defense front against phishing attacks. These directions have great development prospects and research value. The introduction of concept drift detection and incremental learning mechanism enables the model to adapt to the evolution of the attack characteristics in real time, which will greatly enhance the adaptability of the algorithm to face new types of attacks and the effectiveness of defense.

Funding

This work was supported by Beijing Troy Cloud Data Technology Co., Ltd.

References

[1] Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. Computers & Security, 68, 160-196.
[2] Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2017). Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications, 28, 3629-3654.
[3] Vayansky, I., & Kumar, S. (2018). Phishing–challenges and solutions. Computer Fraud & Security, 2018(1), 15-20.
[4] Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or not phishing? A survey on the detection of phishing websites. IEEE Access, 11, 18499-18519.
[5] Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), 2091-2121.
[6] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.
[7] Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). The design of phishing studies: Challenges for researchers. Computers & Security, 52, 194-206.
[8] Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications, 106, 1-20.
[9] Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., & Gupta, C. (2024). Anti-phishing: A comprehensive perspective. Expert Systems with Applications, 238, 122199.
[10] Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.
[11] De Ville, B. (2013). Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448-455.
[12] Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 237-269.
[13] Gavankar, S. S., & Sawarkar, S. D. (2017, April). Eager decision tree. In 2017 2nd International Conference for Convergence in Technology (I2CT) (pp. 837-840). IEEE.
[14] Rana, K. K. (2014). A survey on decision tree algorithm for classification. International journal of Engineering development and research, 2(1), 1-5.

[15] Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences, 12(3), 246-269.
[16] Ahmed, D. S., Hussein, K. Q., & Allah, H. A. A. A. (2022). Phishing websites detection model based on decision tree algorithm and best feature selection method. Turkish Journal of Computer and Mathematics Education, 13(1), 100-107.
[17] Yang, X., Yan, L., Yang, B., & LI, Y. F. (2017, April). Phishing website detection using C4. 5 decision tree. In International conference on information technology and management engineering (ITME 2017) (pp. 119-124).
[18] Machado, L., & Gadge, J. (2017, August). Phishing sites detection based on C4. 5 decision tree algorithm. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-5). IEEE.
[19] Ogonji, D. E. O., Wilson, C., & Mwangi, W. (2023). A hybrid model for detecting phishing attack using recommedation decision trees. In ITM Web of Conferences (Vol. 57, p. 01018). EDP Sciences.
[20] Mithra Raj, M., & Arul Jothi, J. A. (2022, October). Website phishing detection using machine learning classification algorithms. In International Conference on Applied Informatics (pp. 219-233). Cham: Springer International Publishing.
[21] Espinoza, B., Simba, J., Fuertes, W., Benavides, E., Andrade, R., & Toulkeridis, T. (2019, December). Phishing attack detection: A solution based on the typical machine learning modeling cycle. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 202-207). IEEE.