NETWORK AND APPLICATION SECURITY

DOI 10.20535/2411-1031.2024.12.2.315739
UDC 621.618:519

OLEKSANDR ILIASHOV,
SERHII SHOLOKHOV,
OLEKSII KHAKHLIUK,
PAVLO RYZHUK

PROBLEM FORMULATION AND SYNTHESIS OF STATISTICAL ALGORITHMS FOR RECOGNISING WEB RESOURCES AND THEIR VULNERABILITIES BY SIGNATURES OF STATISTICAL AND FUZZY LINGUISTIC FEATURES IN CYBERINTELLIGENCE COMPLEXES

This study addresses the challenge of automating vulnerability recognition in web resources using statistical and fuzzy linguistic features. It presents a formalized approach for the fuzzy recognition of web resource vulnerabilities based on complex reference descriptions defined by signature intervals of statistical and fuzzy feature values. The research introduces algorithms for both single- and multi-alternative recognition of web resources, utilizing decision-making methods such as the minimax rule, Bayesian risk, maximum a posteriori probability, and maximum likelihood.

The primary objective is to enhance the accuracy of vulnerability detection in web resources, especially under conditions of limited training data and fuzzy feature descriptions. The proposed algorithms aim to minimize decision errors and effectively classify vulnerabilities despite uncertain prior probabilities. This is particularly relevant in cybersecurity, where accurate threat detection and classification are critical.

The research also highlights the practical value of these algorithms in improving the efficiency of cyber intelligence systems (CIs) for detecting security breaches and classifying web resource vulnerabilities. The proposed algorithms are designed to adapt to the complex and uncertain nature of web resource security, enabling better analysis of attack scenarios and the development of targeted protection strategies.

In addition, the study identifies several challenges, including the complexity of formalizing reference descriptions for fuzzy features and the difficulties in applying traditional statistical recognition methods to web resources with fuzzy linguistic variables. The paper suggests future research directions, including developing new methodologies for processing large volumes of data and integrating these algorithms into modern cybersecurity systems.

Overall, this research contributes to the field of cyber intelligence by offering novel solutions for automating the detection of web resource vulnerabilities, thus enhancing the security of online systems.

Keywords: statistical recognition, vulnerabilities of web resources, minimax rule, automated recognition, Bayesian criterion, cyber intelligence, automated complexes.

Formulation of the problem in general terms. The present time is characterised by the development of cyber intelligence (CI) systems (Fig. 1) with built-in automated recognition procedures (ARP) for vulnerabilities of Web-resource software for unauthorised access to intelligence information.

The most complex and key tasks to be solved in the automated recognition of Web resource vulnerabilities (possible attack scenarios for unauthorised intrusion) are the formation of a working dictionary of intelligence features and signatures [1], formalisation of the problem and synthesis of algorithms for deciding whether one of the given sample images corresponds to the values of intelligence signatures.

Depending on the mode of operation of the CI complex, the recognition process can use methods for testing simple (CI object search mode) or complex hypotheses (mode of additional investigation of a specific CI object for unauthorised information acquisition), Fig. 1.

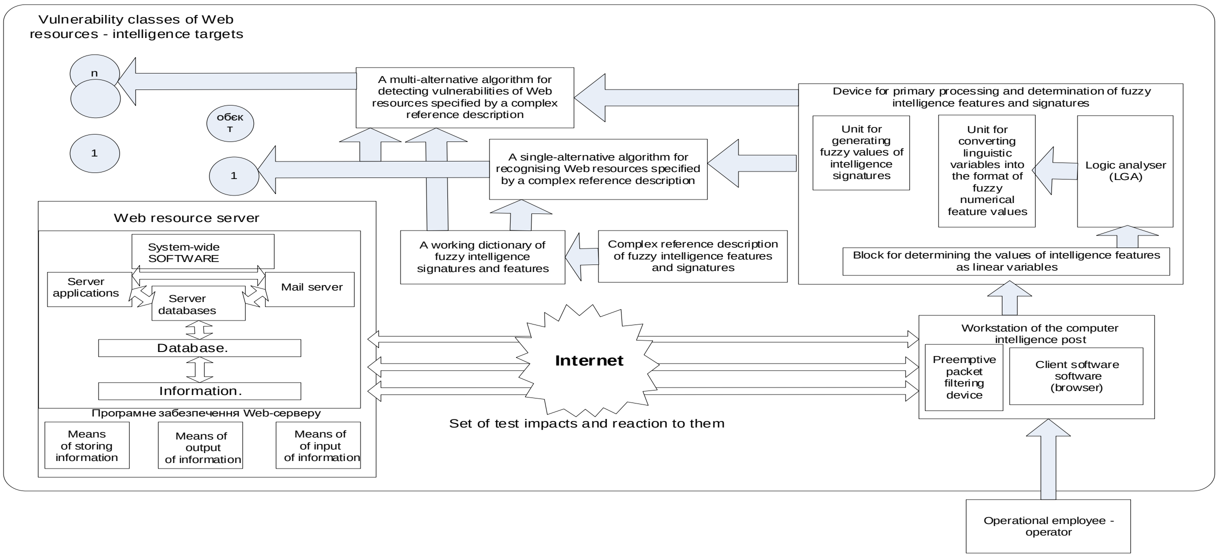
Vulnerability classes of Web resources - intelligence targets
n
обєк т
1
1
A multi-alternative algorithm for detecting vulnerabilities of Web resources specified by a complex reference description
A single-alternative algorithm for recognising Web resources specified by a complex reference description
Device for primary processing and determination of fuzzy intelligence features and signatures
Unit for generating fuzzy values of intelligence signatures
Unit for converting linguistic variables into the format of fuzzy numerical feature values
Logic analyser (LGA)
Block for determining the values of intelligence features as linear variables
Web resource server
System-wide SOFTWARE
Server applications
Mail server
Server databases
Database.
Information.
Програмне забезпечення Web-серверу
Means of storing information
Means of output of information
Means of input of information
A working dictionary of fuzzy intelligence signatures and features
Complex reference description of fuzzy intelligence features and signatures
Internet
Set of test impacts and reaction to them
Workstation of the computer intelligence post
Preemptive packet filtering device
Client software software (browser)
Operational employee - operator

Figure 1 – Block diagram of the unit for obtaining intelligence features and recognition of the automated complex for cyber intelligence of Web resources

At the same time, at the stage of searching for CD objects, one hypothesis is tested against one alternative (the task of classifying the type of object – not a CD object (Fig. 1, block 4)), and during the post-exploration, a set of hypotheses (the task of classifying by type of vulnerability and corresponding attack scenarios for unauthorised penetration, Fig. 1, block 5.

The solution of recognition tasks in CI complexes is complicated by the complexity of formalising signatures and features, their heterogeneous structure, and the presence of overlapping classes of images. This is due to the fact that the input of the device for primary processing and formation of intelligence signatures (Fig. 1) is the results of the application of test effects, which structurally represent the linguistic expressions of the operator of the CI complex such as “true”, “false”, “absolutely true”, “absolutely false”, which are standard for the theory of fuzzy logic. Each value of a linguistic variable is a fuzzy set of a unit interval. Based on the concept of a linguistic variable introduced in [4]–[6], a simple linguistic variable is assigned to each test influence from the set of feature signatures, and a complex linguistic variable to each signature. However, in the literature known to the authors, the issue of formalising the reference description of intelligence signatures and features, synthesising fuzzy algorithms for recognising the vulnerability of Web resources by implementing a set of test impacts – linguistic statements of the operator in hardware and software cyber intelligence systems is not fully resolved.

Analysis of the latest research and publications. The formalisation of a complex reference description for recognition is carried out on the basis of introducing the probabilities of an image belonging to each of the classes of images [2] and their probabilistic and statistical models. However, the fuzzy nature of intelligence features and synatures and the incompleteness of the training sample significantly complicate the development of a reference description and the use of probabilistic and statistical models to determine the security status of Web resources.

This makes it important to consider logical and fuzzy methods and approaches, in the development of which methods and models from the field of soft computing become convenient.

Approaches to the synthesis of fuzzy recognition algorithms have been studied in [4], [5]. However, the results obtained are not adapted to solve CI problems. The issues of forming a reference description of fuzzy signatures for the synthesis of decision-making algorithms for the case of input features obtained in the form of linguistic variables are not defined.

Therefore, the topic of the article is relevant and has practical significance for the practice of intelligence units.

Formulating the objectives of the article. Setting the task and, on its basis, developing a formalised complex reference description of features and an automated fuzzy algorithm for recognising CI objects and vulnerabilities in complexes of CI Web resources.

Presentation of the main research material. The task of classification of CI objects and vulnerabilities (ways of security breaches) of Web resources can be formalised as a fuzzy modification of the pattern recognition task by a sample of linguistic variables.

Task statement. Suppose that the output of the device for primary processing and determining fuzzy intelligence signatures of the CI complex, Fig. 1, records input observations-responses that represent the reaction of a Web resource to a set of test impacts (Fig. 1). The results of the test impacts are determined by the operator in the form of logical fuzzy variables such as “strictly impossible” (SIO), “impossible” (IM), “most likely possible” (MP), “possible” (M) and “strictly possible” (SP).

In aggregate ψ K classes (images) of recognition objects are defined ψ_i, $i \in \{1,2,...,K\}$ CI objects or the state of vulnerability of a Web resource (cyber intelligence objects, types of attacks for unauthorised penetration, etc.)

According to the approach [3], a working dictionary of signatures containing equivalent values is formed that is optimal for a certain indicator, for example, the minimum dimension while ensuring the required level of information content $\Im$ unclear intelligence indications s_i, $i \in \{1,2,...,\Im\}$. Then, each of the K classes of objects or vulnerabilities of the CI Web resources can be characterised by a corresponding signature (vector) $x_i^s = \left\{x_{i1}^{¢}...x_{i1}^{2}, x_{i2}^{¢}...x_{i2}^{2}, ..., x_{i\text{Á}}^{¢}...x_{i\text{Á}}^{2}\right\}, i = 1...K$, $x_i^s \text{ Î } X^s$, $X^s \text{ Ì } R^n$, which essentially represents an ordered set of intervals of fuzzy values of complex linguistic features defined in digital form, Fig. 2. The boundaries of the intervals of fuzzy values of linguistic features are determined by the type of functions $\text{m}_{i}, i = 1...\text{Á}$ in the functioning of the complex of CI Web-resources of the learning mode, Fig. 1, and have the corresponding randomness in determining. A training set N is given in the form of a set of correspondences of the type $(s_i \to x_i)$, Fig. 2. Membership functions μ_j, Fig. 2, allow us to determine the equivalent values of fuzzy linguistic variables in the range of numerical values $[0+i-1...1+i-1]$. Each value of a linguistic feature in an intelligence signature ψ_i the image is assigned to an interval of values (by the number of classes) using the appropriate membership functions μ_i, Fig. 2.

Based on the results of testing the Web resource, the operator generates a sample of certain equivalent values $x = (x_1, x_2, ..., x_\Im)$ linguistic variables resulting from the evaluation of the values of fuzzy features s_i, $i \in \{1,2,...,\Im\}$.

In the ideal case, in the absence of interfering random factors and operator errors, if the test effect of the query-response is positive, it is decided that the value of the linguistic variable is identified with a logical one, Fig. 2, and otherwise with a logical zero. However, in the practice of conducting CI, the accuracy of the interpretation of the test query results is affected by the difficulty in determining the limits of verisimilitude of linguistic variables and a set of random factors. Therefore, the reference description of intelligence signatures is a set of value intervals s_i features with a range from $[0+i-1]$ to $[1+i-1]$, Fig. 2, which are given as a set of corresponding conditional probability densities of feature values $w_{jr}(x^e{}_j, x^{e}{}'_{jr}, x^{e}{}''_{jr})$. These densities characterise the reference distribution of the values of the *j*-th feature $x^e{}_j$ on each *r*-th interval of possible values $x^{e}{}'_{jr}, ..., x^{e}{}''_{jr}$ features for each of the recognised images.

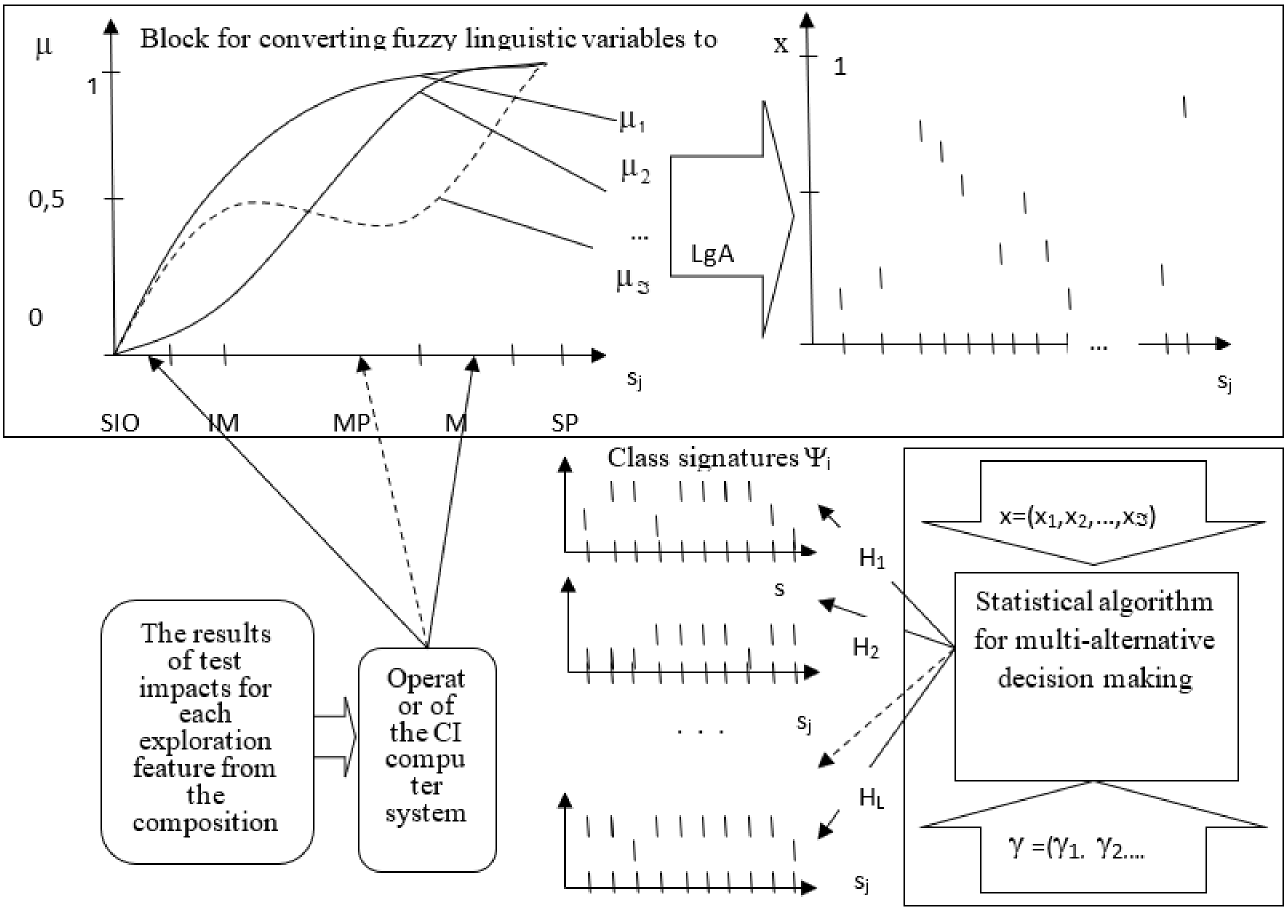
μ
Block for converting fuzzy linguistic variables to
1
0,5
0
μ1
μ2
...
μз
LgA
x
1
sj
sj
SIO
IM
MP
M
SP
Class signatures Ψi
H1
s
H2
sj
. . .
HL
sj
x=(x1,x2,...,xз)
Statistical algorithm for multi-alternative decision making
γ =(γ1. γ2....
The results of test impacts for each exploration feature from the composition
Operat or of the CI compu ter system

Figure 2 – Explanation of the formalisation of the problem of automated recognition of web resource vulnerabilities using fuzzy signatures in cyber intelligence complexes

When applying the recognition algorithms (Figs. 1, 2), the following hypotheses are tested H_i that the sample of ordered values of linguistic fuzzy variables belongs to the image ψ_i. As a result of the inspection, one of L decisions should be made γ_i, that is in the process of accepting a hypothesis H_i.

You need to choose a non-randomised rule δ_L, which allows, in accordance with the optimality criterion, to develop an algorithm that, depending on the results of determining the linguistic feature s on L areas $x_i, i = 1...Á$, that do not intersect, Fig. 2. If during the recognition of the sample value $x = (x_1, x_2, ..., x_з)$ will simultaneously fall into the corresponding areas xi of the signature, a decision is made γ_i.

The foundation for developing decision-making strategies within recognition subsystems of cyber intelligence (CI) complexes lies in the application of statistical algorithms for pattern recognition, as established in studies [4]–[6]. These studies provide a robust methodological framework for designing algorithms aimed at testing complex statistical hypotheses in the context of pattern recognition. These patterns are characterized by a complex reference description, which is commonly encountered in radio control complexes. The algorithms derived from these studies utilize a structured approach to represent and process intricate patterns, enabling effective classification and recognition tasks.

However, it is crucial to note that while the theoretical results from these studies are comprehensive, they remain general in nature and do not address specific applications, such as the recognition of vulnerabilities in Web resources through the use of intelligence signatures. Addressing this gap requires tailoring the foundational methodologies to suit the unique characteristics of Web-based environments. In light of this, the current research builds upon the principles outlined in [4]–[6] and focuses on the synthesis of statistical pattern recognition algorithms that are defined by complex reference descriptions. These descriptions are articulated in the form of signatures representing equivalent values of linguistic features. The goal is to enable their practical application in the CI subsystems of Web resources, thereby advancing the capability to detect and classify vulnerabilities effectively.

This research not only extends the theoretical framework but also bridges the gap between general statistical recognition methodologies and their specific application in the domain of Web resource vulnerability analysis. By leveraging linguistic feature equivalence as a foundational concept, the proposed approach facilitates the software implementation of these algorithms, optimizing their utility in CI complexes dedicated to Web resource security.

Synthesis of algorithms for pattern recognition given a complex reference description in the form of intervals of feature values. In practice, when solving the problems of CI of Web resources, the use of classical decision-making algorithms is not always possible [1]–[2]. This is due to the fact that each of the images in the feature metric (parameters of the testing process) is specified by complex reference descriptions in the form of sets of continuous intervals of reference values of features.

In heuristic algorithms for recognising such patterns, continuous intervals are usually described by sets of discrete values. After that, the resulting discrete set of reference feature values is obtained, which consists of discrete values and continuous intervals converted to the discrete form. Then a set of simple hypotheses is tested about the correspondence of the observed sample to each of the discrete reference values of the features. For this purpose, a set of partial recognition algorithms is used. Based on the intermediate recognition results, a final decision is made on whether the sample belongs to one of the patterns. When performing such algorithms, a large number of operations leads to unnecessary time spent on making a decision. At the final stage of recognition, heuristic algorithms for making final decisions (voting) are often used. This leads to significant probabilities of false decisions, which can be difficult to estimate a priori. In addition, in practice, there are often cases when discrete reference values of features are not equally likely and/or in continuous intervals, the reference distribution of a feature cannot be described by a uniform law. In these cases, the process of synthesising multi-step algorithms for complexes of CIs is further complicated by the need to determine a separate decision threshold for each partial algorithm. For simplicity, in such cases, partial thresholds are sometimes assumed to be equal to each other. This leads to additional decision errors.

It is possible to avoid the above difficulties in building recognition algorithms by applying methods for testing complex statistical hypotheses [5]. In this case, the cumbersomeness of theoretical statements is compensated by the positive effect achieved by a one-step algorithm that is optimal according to one of the statistical optimality criteria. In addition, these algorithms allow to minimise errors arising from parametric uncertainty about the actual values of the observed image features.

We consider the method of synthesis of statistical algorithms for pattern recognition given a complex reference description for two modes of application of complexes in:

a) single-alternative recognition to decide whether an image (an unknown Web resource) is an object of cyber intelligence under conditions of uncertainty about the a priori probabilities of observing images;

b) multi-alternative recognition to make a decision on the classification of the selected object of cyber intelligence into one of the classes, regarding the vulnerabilities of its software.

Let the set ψ of the recognition objects are specified images $\psi_i \subset \psi$, that are recognised. Images are reactions to test influences (values of relevant linguistic variables) – objects of recognition $y_i = \{y_{in}\}, i \hat{I} \{1, 2, ..., L\}, n \hat{I} N_i$, where N_i – is a set of indices, the number of elements v_i of which is equal to the number of types of Web resource reactions to a set of test impacts containing the signature of the *i*-th image (type or type of Web resource vulnerability). Each image can a priori correspond to several intervals of reference values of features - fuzzy linguistic variables, Fig. 2. Each of the images is defined by its own reference description in the field of X_i $\Im$ - dimensional Euclidean reference space X with coordinate axes $x_1, x_2, ..., x_j, ..., x_{\text{Á}}$.

The reference description of the images represents the a priori probability densities of the mixed type obtained as a result of training the recognition algorithm and combined into intelligence signatures $w_i(x)$ reference values of the feature *x* on the set that can be determined as a sum of sum:

$$w_j(x) = \overset{R_j}{\underset{r=1}{\text{å}}} p_{jr} w_{ir}(x^e{}_j, x^e{}'_{jr}, x^e{}''_{jr}), \quad \overset{R_j}{\underset{r=1}{\text{å}}} p_{jr} = 1. \tag{1}$$

$R_{(.)j}$ weighted probability densities $w_{jr}(x^e{}_j, x^e{}'_{jr}, x^e{}''_{jr})$ reference distribution x^e features on each *r*-th interval of possible values $x^e{}'_{jr}, ..., x^e{}''_{jr}$ *j*-th – for each of the recognised images, where are the conditional a priori probabilities of observing the *r*-th interval, $r\ \text{Î}\ \{1, 2,..., R_{(.)j}\}$ *j*-th features in the *i*-th image, $R_{ij} = v_i, R_{qj} = v_q, "j\ \text{Î}\ \{1, 2, ..., \text{Á}\}$.

One reference interval for the value of the *j*-th feature x_j^e, $j = 1, 2,..., \text{Á}$, in one image defines one recognition object (the result of the response to the test influence) in the coordinate of this feature. In this case, the density of a complex reference description is the fulfilment of the following condition:

$$w_{ir}(x^e{}_j, x^{e\prime}{}_{jr}, x^{e\prime\prime}{}_{jr}) = \begin{cases} 0 \text{ for } x^e{}_j < 0 + j - 1, \\ 0...1 \text{ for } x^{e\prime}{}_{jr}\ \text{£}\ x^e{}_j\ \text{£}\ x^{e\prime\prime}{}_{jr}. \\ 0, \text{ for } x^e{}_j > 1 + j - 1. \end{cases} \tag{2}$$

L hypotheses are put forward $H_1, H_2,..., H_L$ that the observed sample x, (size $\zeta \times \Im$) ζ – multiple of the estimated values $\Im$ features belong to one of the images ψ_i, described (corresponds to the values of the feature signature of the corresponding class). The decision space consists of L elements γ_i – *j* decisions to accept a hypothesis H_i.

The task is to choose a non-randomised discrete-analogue rule δ, which implements the division of space X on L areas x_i, $\bigcup_{i=1}^{L} x_i = X$, that do not overlap. If during the recognition process, each feature value from the sample $x = (x_1, x_2, ..., x_{\text{Á}})$ falls within the appropriate signature intervals

$$x_i^s = \left\{x_{i1}\text{¢}...x_{i1}^{\ 2},\ x_{i2}\text{¢}...x_{i2}^{\ 2}, ..., x_{i\text{Á}}\text{¢}...x_{i\text{Á}}^{\ 2}\right\}, i = 1, ..., \mathcal{K} \tag{3}$$

a decision is made γ_i.

In the synthesis of statistical algorithms for recognising Web resources as CI objects (assigning a Web resource to the “class of CI object” or “not a CI object”) that implement methods for testing simple hypotheses, it is difficult to determine the value of the a priori probabilities of observing patterns. In such conditions, in most cases, the minmax criterion is used [5], [6].

Let us consider the synthesis of optimal min-max algorithms for recognising Web resources – CI objects specified by signatures in the form of a complex reference description of type .

It is easy to show that the multi-alternative recognition problem can be easily simplified to the single-alternative case (case “a”) by considering the set of ψ recognition objects on which two images are set $\psi_o \subset \psi$ and $\psi_{no} \subset \psi$ (a CI object or not a CI object). Two hypotheses are tested H_o and H_{no} that the observed sample, $x^{\varsigma\Im}$, size $\zeta \times \Im$, ζ – multiples of measured values of equivalent values $\Im$ of the linguistic features of s_i corresponds to the image signature ψ_o or ψ_{no}. The audit should result in one of two decisions γ_o or γ_{no}, in the adoption of hypotheses H_o or H_{no}, accordingly.

Let us specify the reference distribution of the *j*-th feature s_i, $j = \{1, 2, ..., \text{Á}\}$ (1, 2), as the sum of contingent distributions $w_{ij}(x_j^e) = W(x_j^e \mid \psi_o)$ $w_{qj}(x_j^e) = W(x_j^e \mid \psi_{no})$, weighted with unknown a priori probabilities p_o and $p_{no} = 1 - p_i$ pattern observation p_o and S_{no} [5], [6]

$$W_{ij}(x^e{}_j) = p_i \cdot w_{ij}(x^e{}_j) + p_q \cdot w_{qj}(x^e{}_j) \qquad (4)$$

Each of the probability densities included in (1) $w_{(.)j}(x_j^e)$ represent by weighted sums:

$$\begin{gathered} w_{ij}(x^e{}_j) = \sum_{r=1}^{R_{ij}} p_{ijr} w_{ijr}(x^e{}_j, x^{e\prime}{}_{ijr}, x^{e\prime\prime}{}_{ijr}), \quad \sum_{r=1}^{R_{ij}} p_{ijr} = 1 \\ w_{qj}(x^e{}_j) = \sum_{r=1}^{R_{qj}} p_{qjr} w_{qjr}(x^e{}_j, x^{e\prime}{}_{qjr}, x^{e\prime\prime}{}_{qjr}), \quad \sum_{r=1}^{R_{qj}} p_{qjr} = 1 \end{gathered}, \qquad (5)$$

imposing on (5) the fulfilment of condition (3).

If the probabilities p_{ijr}, p_{qjr} ,are unknown, it is difficult to obtain an unambiguous solution to the problem. The way out of this situation is to initially consider these probabilities equal to each other for each feature in one image. Then, in the course of recognition, they are evaluated and the decision rule is adjusted.

Applying the approach of [4]–[6], the optimal minimum max rule for pattern recognition can be obtained for models (1)–(3) ψ_i and ψ_q, which involves comparing the likelihood ratio $\Lambda(x^{\zeta\Im})$ with a threshold:

$$\delta_{MM} : \Lambda(x^{\zeta\Im}) \underset{\gamma_i}{\overset{\gamma_q}{\gtrless}} \mu_{MM} c^*, \quad c^* = \frac{\Pi_{iq} - \Pi_{ii}}{\Pi_{qi} - \Pi_{qq}}. \qquad (6)$$

Attitude $\mu_{MM} = p_i / p_q$ is at probability p_i, corresponding to the highest value of the Bayesian risk. To do this, following the well-known method of [Levin], it is necessary to equal the average over the regions $X^e{}_i$, $X^e{}_q$ conditional risk functions $r_i(x^s)$, $r_q(x^s)$:

$$\int_{x^e{}_i} r_i(x^e) w_i(x^e) \mathrm{d}x^e = r_i u \int_{x^e{}_q} r_q(x^e) w_q(x^e) \mathrm{d}x^e = r_q \,, \qquad (7)$$

where $r_i(x^e) = \Pi_{ii}[1 - \alpha(x^e)] + \Pi_{iq}\alpha(x^e)$, $r_q(x^e) = \Pi_{qi}\beta(x^e) + \Pi_{qq}[1 - \beta(x^e)]$;

$\text{a}\,(x^e) = \underset{X_q}{\text{ò}} W(x^{\text{zÁ}} \mid x^e \,\text{Î}\, X^e{}_i)\mathbf{d}x^{\text{zÁ}}$;

$\text{b}(x^e) = \underset{X_i}{\text{ò}} W(x^{\text{zÁ}} \mid x^e \,\text{Î}\, X^e{}_q)\mathbf{d}x^{\text{zÁ}}$ – conditional probabilities of first- and second-order errors;

Π^{**}– elements of the loss matrix $\Pi = \begin{Vmatrix} \Pi_{ii} & \Pi_{iq} \\ \Pi_{qi} & \Pi_{qq} \end{Vmatrix}$, $\Pi_{iq} > \Pi_{ii} \ge 0$, $\Pi_{qi} > \Pi_{qq} \ge 0$, whose rows correspond to the hypotheses H_i i H_q, and the columns are the solution γ_i i γ_q;

$W(x^{\text{zÁ}} \mid x^e)$ – sampling likelihood function $x^{\zeta\Im}$ at a fixed vector x^e.

The transcendental equation $r_i = r_q$ in relation to μ is brought to the forefront:

$$\mu_{MM} = \arg\{c^* \alpha(\mu c^*) - \beta(\mu c^*) + c^{**} = 0\}, \qquad (8)$$

where $c^{**} = \dfrac{\Pi_{ii} - \Pi_{qq}}{\Pi_{qi} - \Pi_{qq}}$;

$\alpha(\mu c^*)$, $\beta(\mu c^*)$ – complete (averaged over the respective probability densities $w_i(x^e)$ and $w_q(x^e)$ conditional probabilities of errors of the first and second kind, calculated at the threshold $\mu c*$.

In explicit form, rule (3) is relatively simple to define for the case of independent features, which occurs in most practical applications. In this case, each of the reference descriptions $w_i(x^e)$ and $w_q(x^e)$ is a multiplication $\Im$ the corresponding probability densities (2) $w_{ij}(x^e{}_j)$ аδо $w_{qj}(x^e{}_j)$, the decision is made according to the algorithm:

$$\text{d}_{MM}:\underset{X^e{}_q}{\text{ò}} W(x^{z\text{Á}}\,|\,x^e)\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{qj}}{\text{å}}} p_{qjr}w_{qjr}(x^e{}_j, x^e{}'{}_{qjr}, x^e{}''{}_{qjr})\text{ú}\mathbf{d}x^e\,{}'$$

$$' \left\langle \underset{X^{\text{э}}{}_i}{\text{ò}} W(x^{z\text{Á}}\,|\,x^e)\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{ij}}{\text{å}}} p_{ijr}w_{ijr}(x^e{}_j, x^e{}'{}_{ijr}, x^e{}''{}_{ijr})\text{ú}\mathbf{d}x^e \right\rangle^{-1} \underset{\text{g}_i}{\overset{\text{g}_q}{\underset{<}{>}}} \text{m}_{MM}c*\,, \tag{9}$$

μ_{MM} is still defined by equation (6), in which

$$\text{a}\,(\text{m}c*) = \underset{X^{\text{э}}{}_i}{\text{ò}}\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{ij}}{\text{å}}} p_{ijr}w_{ijr}(x^e{}_j, x^e{}'{}_{ijr}, x^e{}''{}_{ijr})\text{ú}\underset{X_q(\text{m}c*)}{\text{ò}} W(x^{z\text{Á}}\,|\,x^e)\mathbf{d}x^{z\text{Á}}\mathbf{d}x^e, \tag{10}$$

$$\text{b}(\text{m}c*) = \underset{X^e{}_q}{\text{ò}}\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{qj}}{\text{å}}} p_{qjr}w_{qjr}(x^e{}_j, x^e{}'{}_{qjr}, x^e{}''{}_{qjr})\text{ú}\underset{X_i(\text{m}c*)}{\text{ò}} W(x^{z\text{Á}}\,|\,x^e)\mathbf{d}x^{z\text{Á}}\mathbf{d}x^e, \tag{11}$$

$X_i(\mu c*)$ and $X_q(\mu c*)$ – acceptable and critical of the hypothesis H_i areas X_i and X_q, defined at $\Lambda(x^{\zeta\Im}) = \mu c*$.

The quality of the decisions made in accordance with (6) can be assessed by the total conditional error probabilities of the first $\alpha(\mu_{_{MM}}c*)$ and second $\beta(\mu_{_{MM}}c*)$ kind. These probabilities are determined by (8) and (9) at $\mu = \mu_{_{MM}}$.

Extending the methods of synthesis of statistical algorithms [4]–[6] to the multi-alternative case, we introduce the statistics of the *i*-th element of the vector of likelihood ratios when testing complex hypotheses in a species:

$$\text{L}_i(x) = \underset{S_i}{\text{ò}} W(x|\,s)\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{ij}}{\text{å}}} p_{ijr}w_{ijr}(s_j, s'_{ijr}, s''_{ijr})\text{ú}\mathbf{d}s\,'$$

$$' \left\langle \underset{S_1}{\text{ò}} W(x|\,s)\underset{j=1}{\overset{\text{Á}}{\text{Õ}}}\text{ê}\underset{\text{ê}r=1}{\overset{\text{é}R_{1j}}{\text{å}}} p_{1jr}w_{1jr}(s_j, s'_{1jr}, s''_{1jr})\text{ú}\mathbf{d}s \right\rangle^{-1}, \tag{12}$$

where $i = 2, 3, ..., L$.

On the basis of (9), algorithms that are optimal with respect to the main performance criteria can be obtained. To do this, it is necessary to determine the likelihood ratio (9) and compare it with the threshold in accordance with the applied criterion and the conditions of a particular task. In particular, the Bayesian algorithm for multi-alternative pattern recognition has the form:

$$\text{d}_B:\underset{q=2}{\overset{L}{\text{å}}}\left(\text{P}_{qt} - \text{P}_{qi}\right)\frac{p_q}{p_1}\text{L}_q(x) \overset{\text{g}_i}{{}^3}\ \text{P}_{1i} - \text{P}_{1t},\quad t = 1, 2, ..., L,\quad t\,{}^1\ i,\quad i = 2, 3, ..., L \tag{13}$$

where $\Pi_{qi} \geq 0$ – elements of the loss matrix Π of size $L \times L$.

To the area X_i, $i\,\hat{\text{I}}\ \{2, 3, ..., \text{Á}\}$, are the points of the sample space X that satisfy the system of inequalities. Area X_1 is determined from the condition $X_1 = X - \bigcup_{i=2}^{L} X_i$.

When applying the **maximum a posteriori probability criterion**, the decision is made to γ_i, $i = 2, 3, ..., L$, if

$$\delta_{MPC} : p_i \Lambda_i(x) \overset{\gamma_i}{=} \max_{2 \le q \le L} p_q \Lambda_q(x), \quad (p_q / p_1)\Lambda_q(x) \ge 1, \quad q = 2, 3, ..., L, \tag{14}$$

and solutions γ_1, if $(p_q / p_1)\Lambda_q(x) < 1, \forall q \in \{2, 3, ..., L\}$.

When applying the **maximum likelihood criterion**, it is necessary to select the recognition objects in the recognition images that represent them as much as possible and make the most plausible decision regarding these objects. The statistics of the likelihood ratio are checked:

$$\Lambda_i(x) = \frac{\max\limits_{s \in S_i} W(x \mid s)}{\max\limits_{s \in S_1} W(x \mid s)}, \quad S_i = \bigcup_{j=1}^{\Im} \left(\bigcup_{r=1}^{R_{ij}} S_{ijr} \bigcup \bigcup_{d=1}^{D_{ij}} s_{ijd} \right) \quad i = 2, 3, ..., L, \tag{15}$$

where S_{ijr} – r-th intervals of reference values of the i-th image in the feature metric s_j;

s_{ijd} – d-th are the reference values of the i-th image in the metric of the j-th feature.

A decision is made γ_i, $i = 2, 3, ..., L$, if

$$\delta_{MLC} : \Lambda_i(x) \overset{\gamma_i}{=} \max_{2 \le q \le L} \Lambda_q(x), \quad \Lambda_q(x) \ge 1, \quad q = 2, 3, ..., L, \tag{16}$$

and decision γ_1, if $\Lambda_q(x) < 1, \forall q = 2, 3, ..., L$.

Conclusions and prospects for further research. As a result of the conducted research, a formalised description of the problem of fuzzy recognition of cyber intelligence objects and vulnerabilities of Web resources, which are set by a complex reference description in the form of a set of intervals of statistical and fuzzy linguistic features, has been developed. Using the obtained reference description and the mathematical apparatus of the theory of testing complex statistical hypotheses, the results of synthesis of algorithms for single and multi-alternative recognition of Web resources – objects of cyber intelligence by the minimum-maximum decision rule, as well as using the Bayesian, maximum a posteriori probability and maximum likelihood criteria are presented.

The direction of further research is to specify the above decisive rules for cases of application of different models of laws of distribution of a random variable

REFERENCE

[1] S. Tarannum, S.M.M. Hossain, and T. Sayeed, "Cyber Security Issues: Web Attack Investigation" in *Hybrid Intelligent Systems*, vol. 647, Lecture Notes in Networks and Systems, A. Abraham, T.-P. Hong, K. Kotecha, K. Ma, P.M. Mishra, and N. Gandhi, Eds. Cham: Springer, 2023, pp. 1254–1269. doi: https://doi.org/10.1007/978-3-031-27409-1_115.

[2] S. Calzavara, M. Conti, R. Focardi, A. Rabitti, and G. Tolomei, "Machine learning for web vulnerability detection: The case of cross-site request forgery", *IEEE Security & Privacy*, vol. 18, no. 3, pp. 8-16, May – June 2020, doi: https://doi.org/10.1109/MSEC.2019.2961649.

[3] F.G. Veshki, and S.A. Vorobyov, "An efficient approximate method for online convolutional dictionary learning", *arXiv preprint*, Jan. 2023, doi: https://doi.org/10.48550/arXiv.2301.10583.

[4] G.V. Pevtsov, "Synthesis of algorithms for recognizing radio emissions based on the Bayesian rule for testing complex hypotheses", *Radioelectronics and Communications Systems*, vol. 41, no. 4, pp. 49-57, 1998.

[5] G.V. Pevtsov, "Synthesis of algorithms for pattern recognition given complex reference descriptions in the azimuth metric for radio emission sources", *Radioelectronics and Communications Systems*, vol. 43, no. 4, pp. 38-45, 2000.

[6] G.V. Pevtsov, and V.A. Lupandin, "Synthesis of multi-alternative pattern recognition algorithms based on testing complex statistical hypotheses using the maximum a posteriori

probability criterion", *Radioelectronics and Communications Systems,* vol. 44, no. 11, pp. 77-80, 2001.

The article was received 06.09.2024.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] S. Tarannum, S.M.M. Hossain, and T. Sayeed, "Cyber Security Issues: Web Attack Investigation" in *Hybrid Intelligent Systems*, vol. 647, Lecture Notes in Networks and Systems, A. Abraham, T.-P. Hong, K. Kotecha, K. Ma, P.M. Mishra, and N. Gandhi, Eds. Cham: Springer, 2023, pp. 1254–1269. doi: https://doi.org/10.1007/978-3-031-27409-1_115.

[2] S. Calzavara, M. Conti, R. Focardi, A. Rabitti, and G. Tolomei, "Machine learning for web vulnerability detection: The case of cross-site request forgery", *IEEE Security & Privacy,* vol. 18, no. 3, pp. 8-16, May – June 2020, doi: https://doi.org/10.1109/MSEC.2019.2961649.

[3] F.G. Veshki, and S.A. Vorobyov, "An efficient approximate method for online convolutional dictionary learning", *arXiv preprint,* Jan. 2023, doi: https://doi.org/10.48550/arXiv.2301.10583.

[4] G.V. Pevtsov, "Synthesis of algorithms for recognizing radio emissions based on the Bayesian rule for testing complex hypotheses", *Radioelectronics and Communications Systems,* vol. 41, no. 4, pp. 49-57, 1998.

[5] G.V. Pevtsov, "Synthesis of algorithms for pattern recognition given complex reference descriptions in the azimuth metric for radio emission sources", *Radioelectronics and Communications Systems*, vol. 43, no. 4, pp. 38-45, 2000.

[6] G.V. Pevtsov, and V.A. Lupandin, "Synthesis of multi-alternative pattern recognition algorithms based on testing complex statistical hypotheses using the maximum a posteriori probability criterion", *Radioelectronics and Communications Systems,* vol. 44, no. 11, pp. 77-80, 2001.

ОЛЕКСАНДР ІЛЬЯШОВ
СЕРГІЙ ШОЛОХОВ
ОЛЕКСІЙ ХАХЛЮК
ПАВЛО РИЖУК

ПОСТАНОВКА ЗАДАЧІ ТА СИНТЕЗ СТАТИСТИЧНИХ АЛГОРИТМІВ РОЗПІЗНАВАННЯ WEB-РЕСУРСІВ ТА ЇХ УРАЗЛИВОСТЕЙ ПО СІГНАТУРАХ ЗНАЧЕНЬ СТАТИСТЧИНИХ ТА НЕЧІТКИХ ЛІГВІСТИЧНИХ ОЗНАК В КОМПЛЕКСАХ КІБЕРНЕТИЧНОЇ РОЗВІДКИ

Це дослідження присвячене проблемі автоматизованого виявлення вразливостей у програмному забезпеченні веб-ресурсів з використанням статистичних та нечітких лінгвістичних ознак. Описано формалізований підхід до нечіткої ідентифікації вразливостей веб-ресурсів, заснований на складних описах посилок, що визначаються через інтервали підписів статистичних і нечітких значень ознак. У дослідженні запропоновано алгоритми для одноальтернативного та багатоальтернативного розпізнавання веб-ресурсів, що використовують правила прийняття рішень, такі як мінімаксне правило, критерії Баєса, максимуму апостеріорної ймовірності та максимальної правдоподібності.

Основною метою є підвищення точності виявлення вразливостей у веб-ресурсах за умов обмежених навчальних даних та нечітких описів ознак. Запропоновані алгоритми спрямовані на мінімізацію ймовірності помилкових рішень та ефективну класифікацію вразливостей, навіть за невизначених апріорних ймовірностей. Це особливо важливо для кібербезпеки, де точне виявлення загроз і класифікація вразливостей є критичними.

Дослідження також підкреслює практичну значущість цих алгоритмів для підвищення ефективності систем кібернетичної розвідки (КР) у виявленні порушень безпеки та класифікації вразливостей веб-ресурсів. Запропоновані алгоритми розроблені для адаптації до складної та невизначеної природи безпеки веб-ресурсів, що дозволяє ефективніше аналізувати сценарії атак і розробляти стратегії захисту.

Дослідження вказує на низку труднощів, зокрема на складність формалізації описів посилок для нечітких ознак і труднощі застосування традиційних статистичних методів до веб-ресурсів з нечіткими лінгвістичними змінними. Визначено напрямки для подальших досліджень, зокрема, розробку нових методологій для обробки великих обсягів даних та інтеграцію цих алгоритмів до сучасних систем кібербезпеки.

Загалом, це дослідження робить значний внесок у сферу кіберрозвідки, пропонуючи нові рішення для автоматизації виявлення вразливостей веб-ресурсів, що підвищує безпеку онлайн-систем.

Ключові слова: статистичне розпізнавання, уразливості web-ресурсів, мінімаксне правило, байєсовський критерій, комп’ютерна розвідка, автоматизовані комплекси.

Iliashov Oleksandr, doctor of military sciences, full professor, chief researcher of the Military Intelligence Research Institute, Kyiv, Ukraine, ORCID 0000-0002-8099-5057, aleksandr.ilyashov@gmail.com.

Sholokhov Serhii, candidate of technical sciences, associate professor, associate professor of the electronic communications academic department, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine. ORCID 0000-0003-2222-8842, kit.docent71@gmail.com.

Khakhliuk Oleksii, candidate of technical sciences, associate professor of the cybersecurity and information security academic department, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0000-0003-1749-0109, khakhlyuk@gmail.com.

Pavlo Ryzhuk, cadet, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0009-0008-7465-3502, pashok2082.ryuz@gmail.com

Ільяшов Олександр Авксентійович, доктор військових наук, професор, головний науковий співробітник науково-дослідного інституту воєнної розвідки, Київ, Україна.

Шолохов Сергій Миколайович, кандидат технічних наук, доцент, доцент кафедри спеціальних телекомунікаційних систем, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

Хахлюк Олексій Анатолійович, кандидат технічних наук, доцент кафедри кібербезпеки та захисту інформації, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, м. Київ, Україна.

Рижук Павло Сергійович, курсант, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.