UDC 004.056.53(045)

Anna Korchenko, PhD
(National Aviation University)

Formation of linguistic standards for of intrusion detection systems

Method is developed that focuses on the tasks of identifying cyberattacks on computer systems, which is based on mathematical models and methods of fuzzy logic and is implemented through six basic stages. The method allows to improve the process of formalization of linguistic standards receive options to improve the efficiency of construction of the corresponding intrusion detection systems.

Due to the development of digital business and the Internet, malicious software and other cyber threats are becoming more prevalent and pervasive. In this regard, the necessary means to detect cyberattacks on various resources of information systems. For this purpose, special means of combating that is able to remain effective when new types of threats, characterized by unknown or vaguely defined criteria. Apply required methods and models of information security based on fuzzy sets for building detection of anomalies generated by the corresponding offensive environment [1], is the basis for successful response to cyberattacks. Important in detecting anomalies, generated by cyberattacks, is the formation of fuzzy standards [1]. On this basis, the development of methods that improve the process of formalization of receiving linguistic standards of the parameters for the intrusion detection systems, there is actual scientific task.

A number of famous, quite effective developments used to solve these problems, detect cyberattacks, such as: the tuple model to form the basic component for the detection of cyberattacks [1], fuzzy approaches to intrusion detection and detecting anomalies, as well as other developments that are used for solving problems in fuzzy environment [2].

For effective application of known models [1], [3] the formal implementation of the process of formation of fuzzy (linguistic) standards that will allow in a given linguistic variable identifies the search term [4]. On this term by using the corresponding sets of rules to determine the level of abnormal condition created by the impact of the corresponding class of cyberattacks. Based on the analysis of existing research and the relevance of the task the aim of this work is to develop an improved (generalized) method of formation of linguistic standards (MFLS) for intrusion detection systems, operating in formalized fuzzy environment.

For construction of a subset of linguistic standards $\mathbf{T}_{\mathbf{ij}}^{\mathbf{e}}$ (see (13) in [1]), displaying the characteristic judgement of the expert concerning the anomalous state of the parameter P_{ij} we will develop an appropriate method that allows to formalize the process of obtaining standards of the parameters for the specified groups of linguistic variables for a specific environment. Improved MFLS is focused on solving the tasks of identifying cyberattacks on computer systems is a further development of the method of linguistic terms using statistical data [2] and is based on six stages.

Stage 1 – formation of subsets of identifiers linguistic assessments. The creation of the subset $\mathbf{LE_i}$ is based on the set of all possible identifiers (ID) of linguistic evaluations (judgments) expert $\mathbf{LE}$ submitted as

$$\mathbf{LE} = \{\bigcup_{l=1}^{c} LE_l\} = \{LE_1, \ldots, LE_c\}, \ (l = \overline{1,c}), \ (1)$$

and that display used expert judgment to characterize the parameters $\mathbf{P_i}$ [1], [3] when observed in a m-dimensional parametric heterogeneous environment [1], and the c – number of such ID. Next, we form the subset ID of expert judgments

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{LE_i}\} = \{\mathbf{LE_1}, \ldots, \mathbf{LE_n}\}, \ (2)$$

where $\mathbf{LE_i} \subseteq \mathbf{LE}$, $(i = \overline{1,n})$ defined as: $\mathbf{LE_i} = \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{LE_{ij}}\} = \{\mathbf{LE_{i1}}, \ldots, \mathbf{LE_{im_i}}\}$, (3)

in this case $\mathbf{LE_{ij}}$ $(j = \overline{1,m_i})$ ID is a subset of the expert's judgments regarding the values of the parameters P_{ij} (see (8) в [1]) in m-dimensional parametric heterogeneous environment. Taking into consideration (3) we will write formula (2) in the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{LE_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{LE_{ij}}\}\} = \{\{\mathbf{LE_{11}}, \ldots, \mathbf{LE_{1m_1}}\}, \ldots, \{\mathbf{LE_{n1}}, \ldots, \mathbf{LE_{nm_n}}\}\}. \ (4)$$

Thus, taking into account $\mathbf{LE_{ij}} \subseteq \mathbf{LE_i}$, about j-th parameter the expert can apply a set from r_j statements (linguistic estimates), displayed by a subset

$\mathbf{LE_{ij}} = \{\bigcup_{k=1}^{r_j} LE_{ijk}\} = \{LE_{ij1}, \ldots, LE_{ijr_j}\}$, (5), where LE_{ijk} $(k = \overline{1,r_j})$ – k-th identifier of a linguistic assessment of the expert concerning a state j-the parameter when i-th cyberattacks in a certain environment, but r_j – number of identifiers in LE_{ij}.

Further, the expression (4) with (5) takes the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{LE_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{LE_{ij}}\}\} = \{\bigcup_{i=1}^{n} \{\bigcup_{j=1}^{m_i} \{\bigcup_{k=1}^{r_j} LE_{ijk}\}\}\} = \{\{\{LE_{111}, \ldots, LE_{11r_1}\}, \ldots,$$

$$\{LE_{1m_1 1}, \ldots, LE_{1m_1 r_{m_1}}\}\}, \ldots, \{\{LE_{n11}, \ldots, LE_{n1r_1}\}, \ldots, \{LE_{nm_n 1}, \ldots, LE_{nm_n r_{m_n}}\}\}\}. \ (6)$$

It should be noted that the expert Express his opinions on the state of the observed actual values of various parameters in a particular environment, but it can use the same statements from the set of $\mathbf{LE}$, displayed appropriate language identifiers.

Stage 2 – formation of the basic matrix of frequencies. To obtain such matrix is filled in with the set of identifiers of the intervals $\mathbf{N}$ and a subset of these identifiers $\mathbf{N_i}$, which are displayed as $\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{N_i}\} = \{\mathbf{N_1}, \ldots, \mathbf{N_n}\}$, (7)

where $\mathbf{N_i} \subseteq \mathbf{N}$, $(i = \overline{1,n})$ we will define as $\mathbf{N_i} = \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{N_{ij}}\} = \{\mathbf{N_{i1}}, \ldots, \mathbf{N_{im_i}}\}$, (8)

in this case $\mathbf{N_{ij}}$ $(j = \overline{1,m_i})$ – the ID subset of intervals, for determining which linguistic expert carries out the evaluation regarding the values of the parameters P_{ij} (see (8) in [1]) in m-dimensional parametric heterogeneous environment. Taking into account (8) we will write down formula (7) in the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{N_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{N_{ij}}\}\} = \{\{\mathbf{N_{11}}, \ldots, \mathbf{N_{1m_1}}\}, \ldots, \{\mathbf{N_{n1}}, \ldots, \mathbf{N_{nm_n}}\}\}. \quad (9)$$

Further, taking into account $\mathbf{N_{ij}} \subseteq \mathbf{N_i}$, about j-th parameter expert for forming the borders of their assessments may use the set of r_j intervals, the displayed subset $\mathbf{N_{ij}} = \{\bigcup_{k=1}^{r_j} N_{ijk}\} = \{N_{ij1}, \ldots, N_{ijr_j}\}$, (10), where N_{ijk} $(k = \overline{1,r_j})$ – identifier of k-th interval used for the formation of frequencies of occurrence of experts on the current state j-th parameter relative i-th cyberattacks in a certain environment, but r_j – the number of IDs fixed intervals on which the assessment.

Then the expression (9) with (10) takes the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{N_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{N_{ij}}\}\} = \{\bigcup_{i=1}^{n} \{\bigcup_{j=1}^{m_i} \{\bigcup_{k=1}^{r_j} N_{ijk}\}\}\} = \{\{\{N_{111}, \ldots, N_{11r_1}\}, \ldots, \{N_{1m_1 1}, \ldots, N_{1m_1 r_{m_1}}\}\}, \ldots, \{\{N_{n11}, \ldots, N_{n1r_1}\}, \ldots, \{N_{nm_n 1}, \ldots, N_{nm_n r_{m_n}}\}\}\}. \quad (11)$$

Based on the elements of the subsets $\mathbf{LE_{ij}}$ and $\mathbf{N_{ij}}$ formed synthesis table of assessments (table 1), the content of which is based on the current fixation evidence (judgments, evaluations), expert, where f_{ijsq} $(s,q = \overline{1,r_j})$ – elements of empirical data showing the number (frequency) of the same utterances (the use of linguistic assessments of the subset $\mathbf{LE_{ij}}$) expert, characterizing the state of the j-th parameter on the interval ID $N_{ijq} \overset{def}{=} [N_{ijq}^{min}; N_{ijq}^{max}]$ $(q = \overline{1,r_j})$, where N_{ijq}^{min} and N_{ijq}^{max} respectively the lower and upper bound q-th interval.

Table 1

Generalized table of assessments $\mathbf{LE_{ij}}$

$\mathbf{LE_{ij}}$	$\mathbf{N_{ij}}$		
	N_{ij1}	…	N_{ijr_j}
LE_{ij1}	f_{ij11}	…	f_{ij1r_j}
…	…	…	…
LE_{ijr_j}	f_{ijr_j1}	…	$f_{ijr_jr_j}$

Further on the basis of generalized evidence on the elements of the subset LE_{ij} (see table 1) formed the basic matrix of frequencies

$$F_{ij} = \|f_{ijsq}\| = \begin{Vmatrix} f_{ij11} & \ldots & f_{ij1r_j} \\ \ldots & \ldots & \ldots \\ f_{ijr_j1} & \ldots & f_{ijr_jr_j} \end{Vmatrix}. \quad (12)$$

Stage 3 – formation of the derivative matrix of frequencies. To implement this step, you create a vector of sums (VS_{ij}) the appropriate columns of the frequency matrix of (12), i.e.

$$VS_{ij} = \left\| vs_{ijq} \right\| = \left\| vs_{ij1}, ..., vs_{ijr_j} \right\| = \left\| \sum_{s=1}^{r_j} f_{ijs1}, ..., \sum_{s=1}^{r_j} f_{ijsr_j} \right\| = \left\| \bigcup_{q=1}^{r_j} \sum_{s=1}^{r_j} f_{ijsq} \right\|, \ (s, q = \overline{1, r_j}), \quad (13)$$

where f_{ijsq} – are the elements of the matrix F_{ij}. Further by the member of VS_{ij} defined a maximum value according to the formula $vsm_{ij} = \bigvee_{q=1}^{r_j} vs_{ijq}$, (14), which is used to form the derivative matrix of frequencies

$$F_{ij}^{'} = \left\| f_{ijsq}^{'} \right\| = (vsm_{ij} / vs_{ijq}) \left\| f_{ijsq} \right\| <=> F_{ij}^{'} = (vsm_{ij} / vs_{ijq}) F_{ij} = \begin{Vmatrix} f_{ij11}^{'} & \dots & f_{ij1r_j}^{'} \\ \dots & \dots & \dots \\ f_{ijr_j1}^{'} & \dots & f_{ijr_jr_j}^{'} \end{Vmatrix}. \quad (15)$$

Stage 4 – formation of fuzzy terms. The construction of subsets of the fuzzy terms $\mathbf{T_i}$ is based on the set of all possible terms $\mathbf{T}$, showing the specific status of the corresponding parameters from $\mathbf{P_i}$ in m_i-dimensional parametrical sub-environment [1], i.e. $\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{T_i}\} = \{\mathbf{T_1}, \dots, \mathbf{T_n}\}$, (16), where $\mathbf{T_i} \subseteq \mathbf{T}$, $(i = \overline{1, n})$, and $\mathbf{T_i} = \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{T_{ij}}\} = \{\mathbf{T_{i1}}, \dots, \mathbf{T_{im_i}}\}$, (17) thus $\mathbf{T_{ij}}$ $(j = \overline{1, m_i})$ is a fuzzy subset of terms relative to the values of the parameters is a fuzzy subset of terms relative to the values of the parameters P_{ij} (see (8) in [1]). Taking into account (17) formula (16) we will write in the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{T_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{T_{ij}}\}\} = \{\{\mathbf{T_{11}}, \dots, \mathbf{T_{1m_i}}\}, \dots, \{\mathbf{T_{n1}}, \dots, \mathbf{T_{nm_n}}\}\}, \ (j = \overline{1, m_i}). \quad (18)$$

Thus, taking into account $\mathbf{T_{ij}} \subseteq \mathbf{T_i}$ and (18), a subset of the fuzzy terms defined as: $\mathbf{T_{ij}} = \{\bigcup_{s=1}^{r_j} \underset{\sim}{T}_{ijs}\} = \{\underset{\sim}{T}_{ij1}, \dots, \underset{\sim}{T}_{ijr_j}\}$, (19), where $\underset{\sim}{T}_{ijs}$ $(s = \overline{1, r_j})$ – are fuzzy terms, and r_j – the number of members in $\mathbf{T_{ij}}$.

Further, the expression (18) with (19) takes the following form:

$$\{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{T_i}\} = \{\bigcup_{\mathbf{i=1}}^{\mathbf{n}} \{\bigcup_{\mathbf{j=1}}^{\mathbf{m_i}} \mathbf{T_{ij}}\}\} = \{\bigcup_{i=1}^{n} \{\bigcup_{j=1}^{m_i} \{\bigcup_{s=1}^{r_j} \underset{\sim}{T}_{ijs}\}\}\} = \{\{\{\underset{\sim}{T}_{111}, \dots, \underset{\sim}{T}_{11r_1}\}, \dots, \{\underset{\sim}{T}_{1m_11}, \dots, \underset{\sim}{T}_{1m_1r_{m_1}}\}\}, \dots, \{\{\underset{\sim}{T}_{n11}, \dots, \underset{\sim}{T}_{n1r_1}\}, \dots, \{\underset{\sim}{T}_{nm_n1}, \dots, \underset{\sim}{T}_{nm_nr_{m_n}}\}\}\}. \quad (20)$$

Next, it needs to generate the values displayed in the component $\underset{\sim}{T}_{ijs}$, what for the following transformations are used. On $F_{ij}^{'}$ matrix elements according to expression (21) the vector of maxima is under construction $FM_{ij} = \left\| fm_{ijq} \right\| =$

$$\left\| fm_{ij1},..., fm_{ijr_j} \right\| = \left\| \bigvee_{s=1}^{r_j} f_{ijs1}^{'},..., \bigvee_{s=1}^{r_j} f_{ijsr_j}^{'} \right\| = \left\| \bigcup_{q=1}^{r_j} \bigvee_{s=1}^{r_j} f_{ijsq}^{'} \right\|, \ (s,q = \overline{1,r_j}) . \quad (21)$$

Based on FM_{ij} we generate the matrix of membership functions

$$M_{ij} = \left\| \mu_{ijsq} \right\| = \left\| \begin{matrix} \mu_{ij11} & \dots & \mu_{ij1r_j} \\ \dots & \dots & \dots \\ \mu_{ijr_j1} & \dots & \mu_{ijr_jr_j} \end{matrix} \right\|, \quad (22)$$

each element of which is given by the expression $\mu_{ijsq} = f_{ijsq}^{'} / fm_{ijs} \quad (s,q = \overline{1,r_j})$. Using (22), we define the fuzzy sets of terms (numbers) $\underset{\sim}{T}_{ijs}$ on the basis of the expression

$$\underset{\sim}{T}_{ijs} = \{ \bigcup_{q=1}^{r_j} \mu_{ijsq} / x_{ijsq} \} = \{ \mu_{ijs1} / x_{ijs1},..., \mu_{ijsr_j} / x_{ijsr_j} \}, \ (q = \overline{1,r_j}) , \quad (23)$$

where $x_{ijsq} = N_{ijq}^{max} / N_{ijr_j}^{max} \quad (q = \overline{1,r_j})$.

Note that the fuzzy number (FN) $\underset{\sim}{T}_{ijs} \quad (s = \overline{1,r_j})$ accordingly interpreting linguistic utterances of experts $LE_{ijk} \quad (k = \overline{1,r_j})$, the display elements of the subset $\mathbf{LE_{ij}} \subseteq \mathbf{LE}$ (see (6)).

Stage 5 – formation of reference FN. To implement this step we use the fuzzy subset (linguistic) reference standards $\mathbf{T_{ij}^{e}}$ (see (13) in [1]), each of which displays a characteristic judgment of the expert (see step 1) regarding the anomalous state of the parameter P_{ij}. Formation of fuzzy standards is based on the conversion of the corresponding FN (23) of the subset $\mathbf{T_{ij}} \subseteq \mathbf{T}$ and is implemented through three steps. Step 1. The transformation of the fuzzy terms (23) so that for all $\underset{\sim}{T}_{ijs}$ it was fair for the order relation, i.e. $\forall x_{ijsq} : x_{ijsq} < x_{ijsq+1} \quad (q = \overline{1,r_j - 1})$. And also taking into account step 2 and step 3 described in [2].

Stage 6 – visualization of linguistic standards. Implementation of this phase is based on building a geometric image of all reference FN $\underset{\sim}{T}_{ijs}^{e} = \{ \bigcup_{q=1}^{r_{js}} \mu_{ijsq}^{e} / x_{ijsq}^{e} \} =$

$\{\mu^e_{ijs1}/x^e_{ijs1}, \mu^e_{ijs2}/x^e_{ijs2}, \ldots, \mu^e_{ijsr_{js}-1}/x^e_{ijsr_{js}-1}, \mu^e_{ijsr_{js}}/x^e_{ijsr_{js}}\}$, (24), $(q = \overline{1, r_{js}})$ belonging to the subset $\mathbf{T}^e_{ij}$ (see (13) in [1]). The locus of points in the plane is defined by a polyline connecting the points representing the components of FN $\underset{\sim}{T}^e_{ijs}$ in ascending order of their supports (media) x^e_{ijsq}. Visualization of one standard reference term (24) is presented in the broken line form ⟵●⟶ on figure 1.

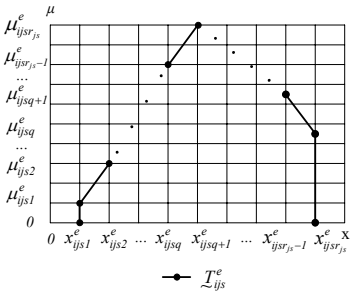

Figure 1. The linguistic standard bass $\underset{\sim}{T}^e_{ijs}$

The proposed improved MFLS for intrusion detection systems, which through the use of sets of identifiers of linguistic assessments and of the identifiers of intervals, basic and derivative matrix frequency display of the judgments of experts regarding cyberattacks characterizing the current state of the settings and processes of formation at given intervals of frequencies of occurrence of expert assessments and subsets in fuzzy terms, allows to formalize the procedure of obtaining reference values of the parameters specified groups of linguistic variables, characterizing in various conditions of the anomalous specific parametric heterogeneous environment.

References

1. Korchenko A.A. The tupel model of basic components' set formation for cyberattacks, Legal, regulatory and metrological support information security system in Ukraine, 2014, V.2 (28), pp. 29-36.
2. Korchenko A.G. The development of information protection systems based on the fuzzy sets, The theory and practical solutions, Kuev, 2006, 320 p.
3. Anna Korchenko, Kornel Warwas, Aleksandra Kłos-Witkowska. The Tupel Model of Basic Components' Set Formation for Cyberattacks // Proceedings of the 2015 IEEE 8th International Conference on «Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications» (IDAACS'2015), Warsaw, Poland, September 24-26, 2015: Vol. 1. – pp. 478-483.
4. Korchenko A.A. The formation method of linguistic standards created for the intrusion detection systems, Zahist ìnformacìï, T.16, №1, 2014, pp. 5-12.