ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

UDC 004.732.056

S.Yu. Gavrylenko, M.S. Melnyk, V. V. Chelak

DEVELOPMENT OF A HEURISTIC ANTIVIRUS SCANNER BASED ON THE FILE'S PE-STRUCTURE ANALYSIS

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

Анотація. Розглянуто методи побудови антивірусних програм, їх переваги та недоліки. Проаналізовано PE-структуру шкідливого та безпечного програмного забезпечення. Знайдено API-функції та строки, притаманні цим файлам та виділено частину із них для подальшого аналізу. Виділені ознаки використано в якості вхідних даних системи нечіткого виведення. Розроблено модель евристичного аналізатора на базі методу нечіткої логіки Мамдані та проведено тестування розробленої системи. Отримані результати досліджень показали можливість використання розробленої системи ідентифікації шкідливого програмного забезпечення в евристичних аналізаторах систем виявлення вторгнень..

Ключові слова: антивірусне програмне забезпечення, комп'ютерна система, шкідливе програмне забезпечення, сигнатурний метод, евристичний метод, PE-структура файлу, нечітка логіка Мамдані.

Аннотация. Рассмотрены методы построения антивирусных программ, их преимущества и недостатки. Проанализировано PE-структуру вредоносного и безопасного программного обеспечения. Найдено API-функции и сроки, присущие этим файлам и выделена часть из них для дальнейшего анализа. Выделенные признаки использовано в качестве входных данных системы нечеткого вывода. Разработана модель эвристического анализатора на базе метода нечеткой логики Мамдани и проведено тестирование разработанной системы. Полученные результаты исследований показали возможность использования разработанной системы идентификации вредоносного программного обеспечения в эвристических анализаторах систем обнаружения вторжений.

Ключевые слова: антивирусное программное обеспечение, компьютерная система, вредоносное программное обеспечение, сигнатурный метод, эвристический метод, PE-структура файла, нечеткая логика Мамдани.

Abstract. Methods for constructing antivirus programs, their advantages and disadvantages are considered. The PE-structure of malicious and secure software is analyzed. The API-functions and strings inherent in these files are found and some of them are selected for further analysis. The selected features are used as inputs for the system of fuzzy inferences. A model of a fuzzy inference system based on the Mamdani fuzzy logic method is developed and tested. The obtained results of the research showed the possibility of using the developed malicious software identification system in heuristic analyzers of intrusion detection systems.

Keywords: antivirus software, computer system, malicious software, signature method, heuristic method, PE-structure of a file, Mamdani fuzzy logic.

Formulation of the task

The first big hacker attack in Ukraine was conducted in 2014 on information systems of the Central Election Commission. In June of 2017 Ukrainian institutions became victims of a bigger cyber attack carried out through a computer virus named "Petya.A". For the first time in the history of Ukraine a hacker attack stopped the work of banks, gas stations, shops and websites of government institutions for several hours. The websites of the Cabinet of Ministers of Ukraine and several of the largest mass media companies were paralyzed. This malicious program has struck computers of many organizations and private individuals in 60 countries around the world. The losses from the virus attack are estimated at $8 billion. [1].

Computer security experts report that the amount of computer viruses and malicious software increases at an alarming rate. Despite all the efforts of researchers and developers in this branch of science, at the moment there is no such antivirus program that could detect all security threats. Therefore, the problem of developing and improving antivirus tools remains an urgent scientific task.

The aim of the article

The aim of the article is to develop a system of fuzzy inference based on the Mamdani fuzzy logic method.

The solution of the problem

Analysis of the literature [2-7] has shown that there are two main operation methods of antivirus programs - the signature method and the heuristic method (fig.1). The signature method is based on scanning and comparison with the standard (mask). The mask contains a set of malicious commands specific to this type of us.

The essence ofheuristic analysis is in the verification of possible habitats of viruses and the detection of commands (groups of commands) that are specific to viruses. Using this method the antivirus monitors all actions that can be implemented by the program. For example, opening or writing to a file, intercepting interrupt vectors, etc. It is in this way that potentially dangerous actions that are specific to viruses are monitored. Through the help of a heuristic code analyzer up to 92% of viruss were found. This mechanism is quite effective but very often leads to false triggering. Files in which the heuristic analyzer has detected a suspicion of a virus are called possibly infected, or suspicious.

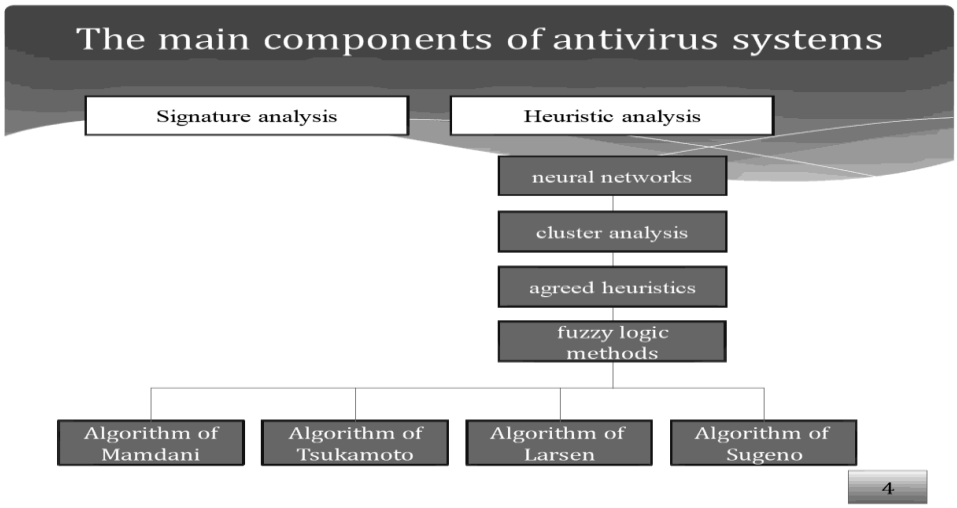

Figure 1 – The main components of antivirus systems

Heuristic methods are based on the use of rules or statistical methods: systems based on weights and rules, cluster analysis, agreed heuristics, expert systems, neural networks, immune networks, etc.

The conducted researches has shown that one of the perspective directions of heuristic analysis of computer viruses is the use of fuzzy logic [4]. Fuzzy logic is used in the analysis of new markets, stock market estimation, assessment of political ratings, optimal price strategy choice, etc.

Fuzzy inference systems are designed to implement the process of fuzzy inference and serve as a conceptual basis for all modern fuzzy logic.

Fuzzy inference is a central point in fuzzy logic and fuzzy control systems [8-10]. The process of fuzzy inference is a procedure or algorithm for obtaining fuzzy conclusions based on fuzzy conditions or assumptions using concepts of fuzzy logic given above. This process unites all the basic concepts of the fuzzy sets theory: membership functions, linguistic variables, fuzzy logical operations, methods of fuzzy implication and fuzzy composition.

The Mamdani algorithm was one of the first to be used in fuzzy inference systems [8]. It was suggested in 1975 by the English mathematician Ebrahim Mamdani as a method for controlling a steam engine. The Mamdani algorithm is shown in fig.2.

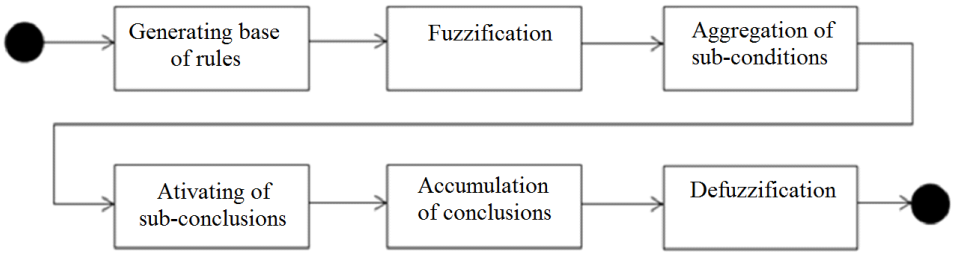

Figure 2 – The process of fuzzy inference by the method of Mamdani

The input data for the computer virus detection system based on the Mamdani fuzzy logic are formed on the basis of the file's PE-structure analysis.

The PE-structure of malicious and secure software was analysed:

- 290 files of type Worm;
- 1050 files of type Trojan;
- 1153 files of type Backdoor;
- 1000 safe files;

and their attributes were identified in the form of API-functions and strings.

The detected attributes were analyzed. 50 most frequently repeated and such, the presence of which in the program code of the file guarantees the belonging of this file to the group of viruses were identified (tab.1).

Rows that were equally common in malicious and in secure files were excluded. Then these strings were separated into groups (A-D) to specify a smaller number of input variables. The groups were identified on the basis of the level of assurance that the files containing these lines were malicious or secure (tab.1).

Table 1 – The input strings

№	Group	Strings	% malicious	% safe	difference
1	A	callnexthookex	27	75	48
		getcurrentprocessid	39	87	48
		getdevicecaps	27	66	39
2	B	getmonitorinfo	19	51	32
		getdesktopwindow	25	54	29
		shellexecute	37	63	26
3	C	getsysteminfo	26	48	22
		unhookwindowshook	28	48	20
		setwindowshook	28	47	19
		regqueryvalue	76	94	18
4	D	wininet	42	22	-20
		gethostbyname	29	3	-26
		getstartupinfo	84	55	-29
		socket	46	16	-30
		regiterserviceprocess	31	0	-31
		inet_addr	35	3	-32
5	E	copyfile	81	37	-44
		wnet	58	8	-50

Input linguistic variables for the system of fuzzy inference by the Mamdani method were described by the following values:

< α, T, X, G, M>, where

α is the name of the linguistic variable(A, B, C, D, E);

T is an array of values (terms) of the input linguistic variable;{ "Danger", "None", "Safe"};

X is an array of values of the input variable (percentage of the number of strings in the file [0; 100]);

G is the procedure of aggregation of conditions (new terms);

M is the function of forming a fuzzy array of values for each term of a set linguistic variable

A trapezoidal function (1) is used as a membership function.

$$MF(x) = \begin{cases} 1 - \dfrac{b-x}{b-a}, & a \le x \le b; \\ 1, & b < x \le c \\ 1 - \dfrac{x-c}{d-c}, & c < x \le d \\ 0, & x \notin (a,d) \end{cases} \quad (1)$$

Specification of the linguistic variable "A" is shown in fig.3.

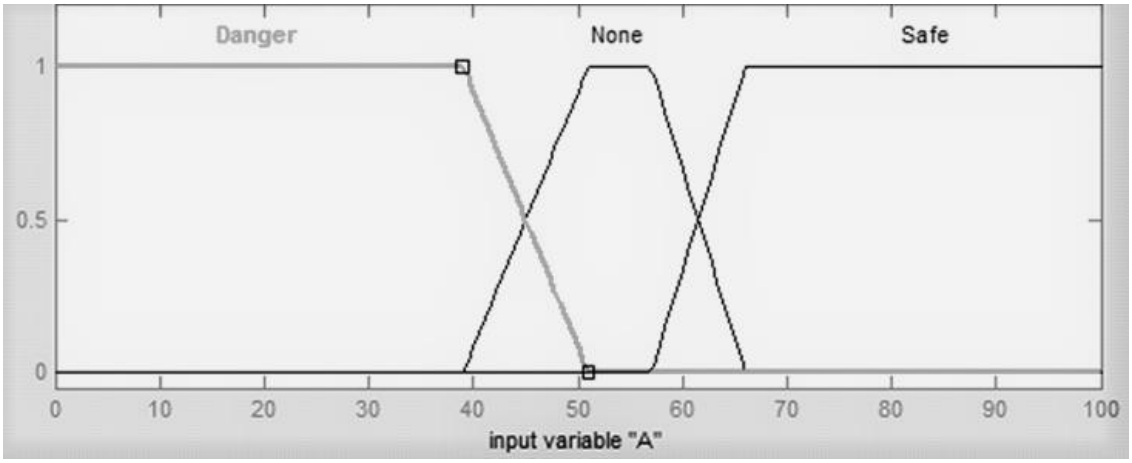

Figure 3 – The graph of the membership function of the input variable "A"

Linguistic variables B, C, D, E are formed according to the table. 1.

The output linguistic variable of the system is specified as "F". The array of values of the linguistic variable "F" is specified by: "Safe", "None", "Virus". A trapezoidal function was used as a membership function (fig.4).

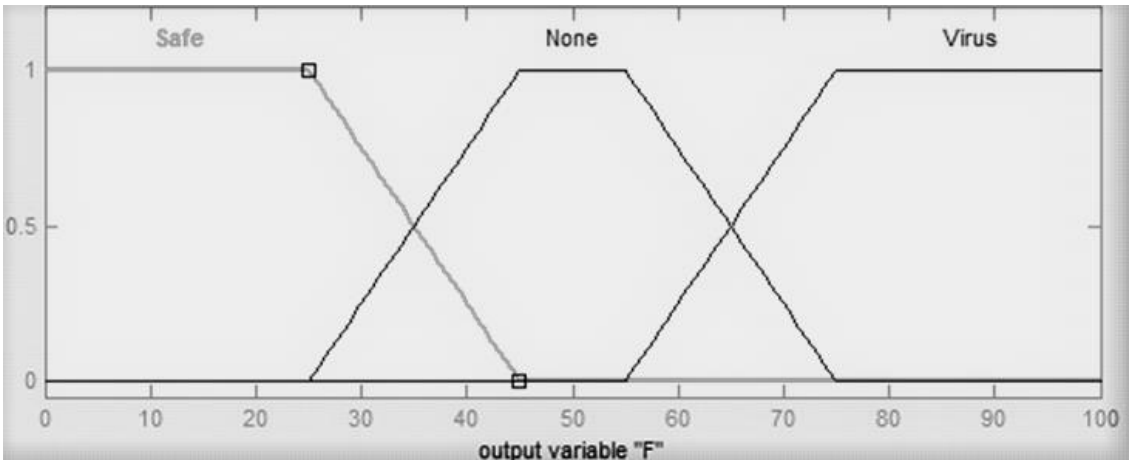

Figure 4 – The graph of the membership function of the output linguistic variable "F"

The base of rules is formed on the basis of prevailing conditions and conclusions, the input and output linguistic variables [11]. Figure 5 shows a fragment of the base of rules for a developed heuristic analyzer based on the Mamdani fuzzy logic method. The base of rules of this analyzer contains 243 different rules.

1. If (A is Safe) and (B is Safe) and (C is Safe) and (D is Safe) and (E is Safe) then (F is Safe) (1)
2. If (A is Danger) and (B is Danger) and (C is Danger) and (D is Danger) and (E is Danger) then (F is Virus) (1)
3. If (A is None) and (B is None) and (C is None) and (D is None) and (E is None) then (F is None) (1)
4. If (A is Danger) and (B is Danger) and (C is Danger) and (D is Danger) and (E is Safe) then (F is Virus) (1)
5. If (A is Danger) and (B is Danger) and (C is Danger) and (D is Safe) and (E is Danger) then (F is Virus) (1)
6. If (A is Danger) and (B is Danger) and (C is Safe) and (D is Danger) and (E is Danger) then (F is Virus) (1)
7. If (A is Danger) and (B is Safe) and (C is Danger) and (D is Danger) and (E is Danger) then (F is Virus) (1)
8. If (A is Safe) and (B is Danger) and (C is Danger) and (D is Danger) and (E is Danger) then (F is Virus) (1)
9. If (A is Safe) and (B is Safe) and (C is Safe) and (D is Safe) and (E is Danger) then (F is Safe) (1)
10. If (A is Safe) and (B is Safe) and (C is Safe) and (D is Danger) and (E is Safe) then (F is Safe) (1)
11. If (A is Safe) and (B is Safe) and (C is Danger) and (D is Safe) and (E is Safe) then (F is Safe) (1)

Figure 5 – A fragment of the base of rules for a fuzzy inference system

Then, to solve a more specific problem, the system of fuzzy inference produces conclusions based on the values of coefficients of each subcondition of the rules, and based on the established base of rules gives a certain fuzzy conclusion. The subjective initial estimate goes through the process of defuzzification, that is the process of transition from the membership function of the output linguistic variable to its certain numerical value [11]. The defuzzification process can be carried out by various methods, for example, by the center of gravity method (2):

$$\Delta = \frac{\int_{\min}^{\max} xMF(x)dx}{\int_{\min}^{\max} MF(x)dx} \tag{2}$$

The simulation results and the visualization are shown in fig..6-8.

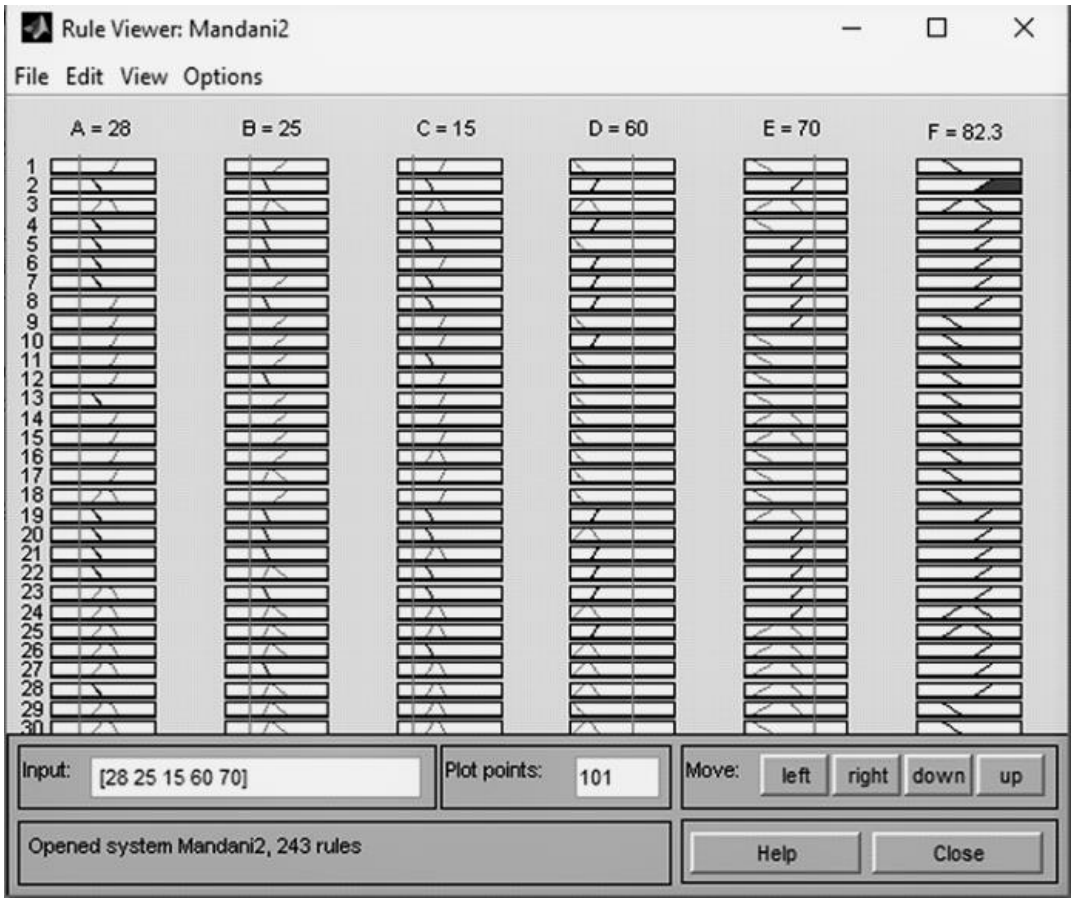

Figure 6 – The simulation results for a malicious file

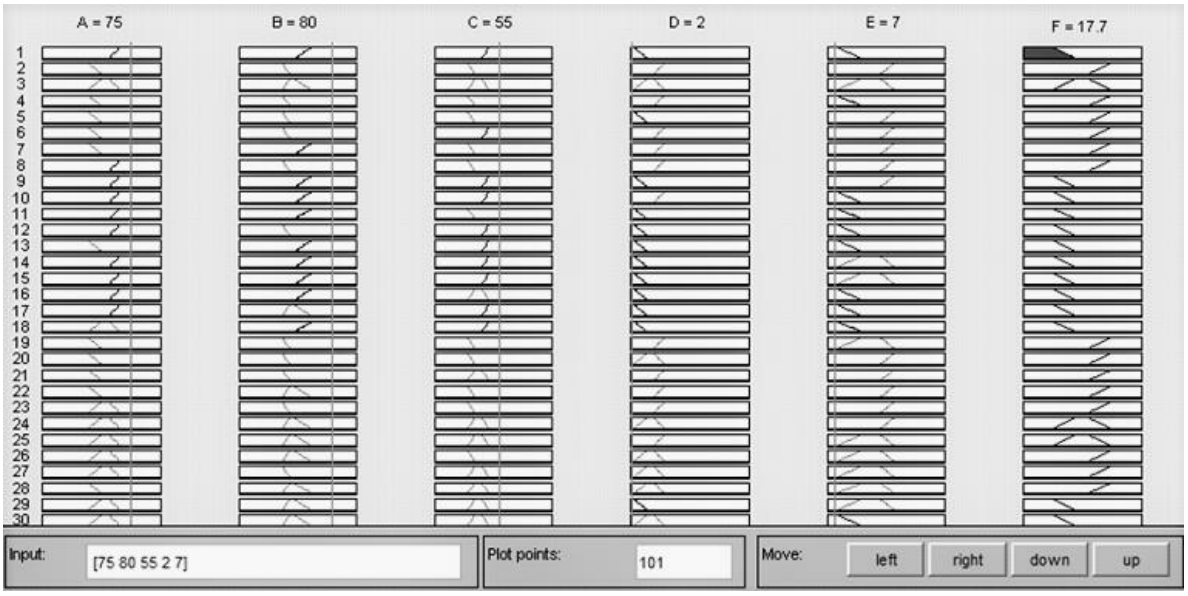

Figure 7– The simulation results for a safe file

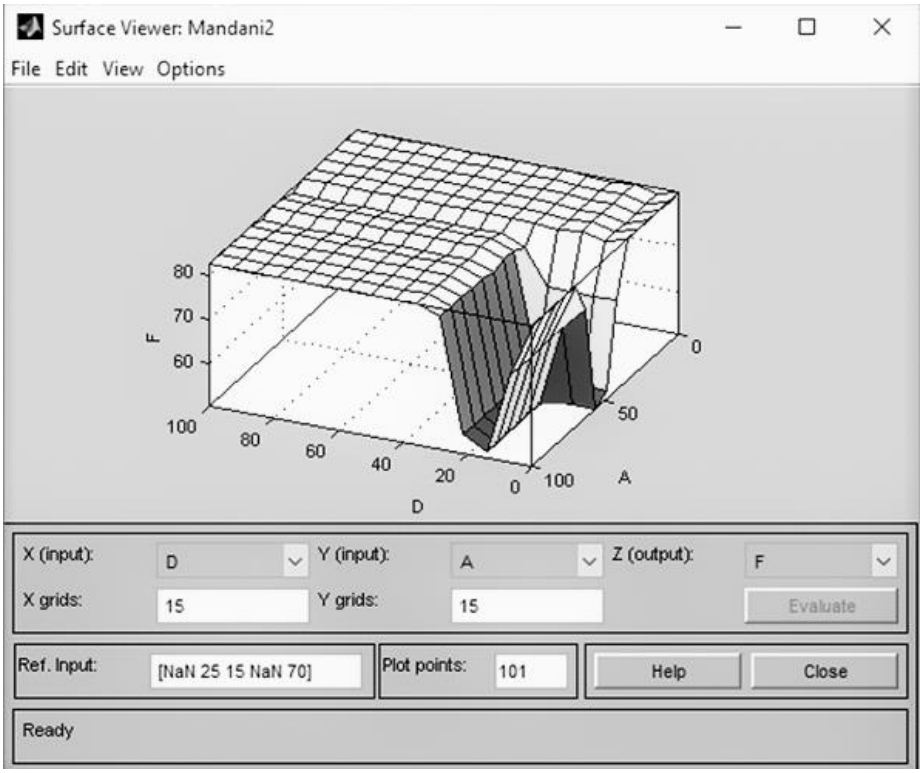

Figure 8 – Visualization of the interdependence between input variables D and A.

As can be seen from the results, the developed heuristic scanner of the anti-virus program based on the file's PE-structure analysis allows us to create multi-level fuzzy productional models, and the used fuzzy output mechanism based on the Mamdani algorithm allows us to obtain the numerical value of the risk of malware detection, the linguistic specification of the risk level and the degree of the expert confidence in the occurrence of this event. The received information will allow us to make a conclusion and develop measures to prevent the infection of the computer system.

Conclusions

1. This article examines methods of creating antivirus programs. The file's PE-structure was analyzed and its attributes were identified in the form of API-functions and strings. Some of the identified attributes were used for analysis.

2. The model of a fuzzy inference system based on the Mamdani fuzzy logic method was developed and tested.

3. The simulation results confirm the possibility of using a heuristic analyzer based on the Mamdani fuzzy logic method as an additional tool for detecting virus attacks in a common system for detecting malicious software.

4. The subjective choice of membership functions and the formation of a base of rules are the disadvantages of this approach.

5. Further improvement of this model can be in the development of a computer system's current events analyzer, which will make the heuristic analyzer more dynamic, capable of working with large volumes of input data and capable to quickly, effectively and qualitatively detect threats for the information security of the computer system.

References

1. Computer security expert estimated the world losses from the virus "Petya" attack. [Web source]. – Access mode: https://tsn.ua/svit/kiberekspert-ociniv-zbitki-vid-virusu-petya-a-u-sviti-953633.html

2. Shelukhin O.I., Sakalema D.Zh., Filinova A.S. Obnaruzheniye vtorzheny v kompyuternye seti (setevye anomalii) [Detection of intrusions into computer networks (network anomalies)] / O.I. Shelukhin, D. Zh. Sakalema, A.S. Filinova. – M.: Goryachaya Liniya-Telekom, 2013. – 220 p.

3. Kaspersky K. Zapiski issledovatelya kompyuternyh virusov [Sketch-book of a computer virus researcher] / K. Kaspersky. – St.P.: Piter, 2006. – 316 p.

4. Zaichenko Yu.P. Nechetkye modeli i metody v intellektualnyh systemah [Fuzzy models and methods in intelligent systems / Yu.P. Zaichenko. – K.: Slovo, 2008. – 344 p.

5. Semenov. S.G. Zashchita dannyh v kompyuterizirovannyh systemah upravleniya (monographiya) [Data protection in computerized control systems (monograph)] / S.G. Semenov, V.V. Davydov, S.Yu. Gavrilenko. – LAP LAMBERT ACADEMIC PUBLISHING GmbH & Co. KG, Germany, 2014.– 236 p.

6. Lukatskiy A.V. Obnaruzheniye atak [Detection of attacks]. – St.P.: BHV-Petersburg, 2001. – 624 p.

7. Lenkov S.V. Metody i sredstva zashchity informatsii [Methods and means of information protection. In 2 vol.] / S.V. Lenkov, D.A. Peregudov, V.A. Khoroshko. – Edited by V.A. Khoroshko. – K.: Ariy, 2008. – Vol. 2. Information security. – 344 p.

8. Kavun S.V. Informatsiyna bezpeka: pidruchnyk [Information security: guide]. – Kharkiv: Edition of KNUE, 2009. – 368 p.

9. Zadeh L. The concept of a linguistic variable and its application to approximate reasoning. – M.: Mir, 1976. – 166 p.

10. Fuzzy sets and probability theory. Recent achievements / Edited by. R.R. Yager. - M.: Radio i svyaz, 1986. – 408 p.

11. Pivkin V.Y., Bakulin E.P., Korenkov D.I. Nechetkiye mnozhestva v sistemah upravleniya [Fuzzy sets in control systems]. – Novosibirsk: Edition of NCU, 1998. – 75 p.

12. Stovba S.D. Proyektirovaniye nechetkih sistem sredstvami MATLAB [Designing of fuzzy systems using MATLAB tools]. – M.: Goryachaya Liniya-Telekom, 2007. – 288 p.

13. Leonenkov A.V. Nechetkoye modelirovaniye v srede MATLAB i fuzzyTECH [Fuzzy modeling in the MATLAB and fuzzyTECH environment]. – St.P.: BHV-Petersburg, 2005. – 736 p.

Article received: 13.11.2017.

Information about the authors

Gavrylenko Svitlana, PhD Tech., Associate Professor, Professor at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str., 2, Kharkiv, Ukraine, 61002, ORCID ID: 0000-0006-4561-8368.

Melnyk Marharyta, Bachelor, Student at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Str. Kyrpychova, 2, Kharkov, Ukraine, 61002, ORCID ID: 0000-0003-0619-7281

Chelack Victor, Student at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Str. Kyrpychova, 2, Kharkov, Ukraine, 61002, ORCID ID: 0000-0001-8810-3394