UDC 651.34 **doi: 10.20998/2522-9052.2020.2.02**

S. Gavrylenko[1], V. Chelak[1], O. Hornostal[1], V. Vassilev[2]

[1] National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, Ukraine
[2] Technical University of Sofia, Sofia, Bulgaria

DEVELOPMENT OF A METHOD FOR IDENTIFYING THE STATE OF A COMPUTER SYSTEM USING FUZZY CLUSTER ANALYSIS

Abstract. The subject of this article is the study of methods for identifying the state of computer systems. **The purpose** of the article is to develop a method for identifying the abnormal state of a computer system based on fuzzy cluster analysis. **Objective**: to analyze methods for identifying the state of computer systems; to conduct research on the selection of source data; to develop a method for identifying the state of a computer system with a small sample or fuzzy source data; to investigate and justify the procedure for comparing fuzzy distances between grouping centers and clustering objects; to develop a software and test. **The methods** used in the paper: cluster analysis, fuzzy logic tools. The following **results** were obtained: a method was theoretically substantiated and investigated for identifying the state of a computer system with a small sample or fuzziness of the initial data, which is distinguished by the use of the method based on fuzzy cluster analysis by the refined grouping procedure. To solve the clustering problem, we used a special procedure for comparing fuzzy distances between grouping centers and clustering objects. Software was developed and testing of the developed method was performed. The quality of classification based on the ROC analysis is assessed. **Conclusions**. The scientific novelty of the results is as follows: a study was conducted on the selection of source data for analysis; a method for identifying the state of a computer system based on fuzzy cluster analysis using a special procedure for comparing fuzzy distances between grouping centers and clustering objects has been developed. This allowed to improve the classification quality to 22%.

Keywords: state identification; computer system; cluster analysis; fuzzy output.

The problem formulation

Identification of the state of a computer system (CS) is one of the ways to rate the quality of its functioning. The state identification system is based on two classes of methods: for identifying abnormality and for identifying abuses [1]. The basis of the methods for identifying abnormality is the construction of patterns for normal CS behavior [2-4]. Abnormality is characterized by data that do not satisfy certain concepts of normal behavior. Methods for identifying abuses are based on the description of known violations, attacks or distortion of relevant information [6]. If the behavior of an object which is similar in description to a known attack is observed, then it is considered that the object was invaded. The most effective methods for identifying anomalies and abuses are machine learning methods. These include classical methods [7, 8], reinforced learning methods [9], ensemble methods [10-12], neural networks and deep learning [13-15]. The basis of these methods is technology and procedures that solve the problem of classifying the state. The analysis showed that the main disadvantages of many of these methods are the neglect of fuzzy data factors and low adaptation to dynamic changes in the structures of the source data and external influences. This, in turn, leads to a decrease in the reliability and efficiency of identifying CS state.

Studies of existing computerized systems for identifying states [16] also revealed a number of limitations on their use. The operation of such systems is based on the detection of network intrusions by comparing the functioning of the system with the profile of their normal operation. However, with the appearance of new abuses and anomalies caused by intrusions with unknown or unclear properties, these systems are not always effective and require long time-consuming for their appropriate adaptation, which also leads to a decrease in the reliability and efficiency of identification [2]. In addition, existing computerized systems, as a rule, are rather expensive, have a closed code and require periodic support of highly qualified specialists to improve them and appropriate settings with the requirements of specific organizations. In this regard, the development of methods and means of identifying the state of the CS is of particular relevance. Moreover, a feature of such methods is to increase the efficiency and reliability of identification in a small sample or fuzzy source data.

Development of a method for classifying the state of a CS based on the development of a fuzzy cluster classifier

One of the effective methods of data classification is the use of the classical method based on cluster analysis. This method is widely used to identify the state of CS [17–19]. But in a real situation, when the selection of the source data is small, when the CS is operating in a critical or non-stationary mode, there is no certainty that random output data are distributed normally. For the same reason, the errors of statistical estimates of mathematical expectations and variance of controlled indicators can be unpredictably large. In this case, the most effective is the use of a fuzzy mathematics apparatus, which is adapted to identify the state of CS under these conditions [20]. At the same time, the use of fuzzy clustering methods (FCM, Fuzzy C-Means), provides that for each element it is necessary to calculate the degree of its belonging to each of the clusters. In this case, there are difficulties in comparing fuzzy numbers. In practice, various heuristic approaches are used to solve the problem of comparing fuzzy numbers. Such approaches are difficult to implement [21]; a specific comparison result can be obtained only in the case of an obvious advantage of one number over another (for example, if there are no intersections of membership functions of the comparing numbers).

We proposed [22] the following simpler and more reliable procedure for comparing fuzzy numbers, based on comparing a fuzzy function of the distances' difference with zero. Let the fuzzy output data be x and y for two states H1 and H2 given by the following membership functions (1):

$$\mu(x)=\begin{cases}0, & x\le b_x;\\ (x-b_x)/(m_x-b_x), & b_x<x<m_x;\\ (c_x-x)/(c_x-m_x), & m_x<x\le c_x;\\ 0, & x>c_x;\end{cases}$$

$$\mu(y)=\begin{cases}0, & y\le b_y;\\ (y-b_y)/(m_y-b_y), & b_x<y<m_y;\\ (c_y-y)/(c_y-m_y), & m_y<y\le c_y;\\ 0, & y>c_y;\end{cases} \qquad (1)$$

$$b_x=\min\{x_{1,}^{(1)}x_{2,}^{(1)}...,x_{l_1}^{(1)}\};\ c_x=\max\{x_{1,}^{(1)}x_{2,}^{(1)}...,x_{l_2}^{(1)}\},$$

$$b_y=\min\{x_{1,}^{(2)}x_{2,}^{(2)}...,x_{l_1}^{(2)}\};\ c_y=\max\{x_{1,}^{(2)}x_{2,}^{(2)}...,x_{l_2}^{(2)}\},$$

$$m_x=\frac{1}{l_1}\sum_{s=1}^{l_1}X_s^{(1)};\quad m_y=\frac{1}{l_2}\sum_{s=1}^{l_2}X_s^{(2)}.$$

Let's find the membership function (2) of their difference $z=x-y$:

$$\mu(z)=\begin{cases}0, & y\le b_z;\\ (x-b_z)/(m_z-b_z), & b_z<z<m_z;\\ (c_y-z)/(c_z-m_z), & m_z<z\le c_z;\\ 0, & z>c_z;\end{cases} \qquad (2)$$

$$b_z=b_x-b_{y,}\quad m_z=m_x-m_{y,}\quad c_z=c_x-c_y.$$

Now the original problem of comparing *x* and *y* is reduced to a simpler problem, namely, comparing a fuzzy number *z* with zero.

Let's introduce the rules for interpreting the result of comparing a fuzzy *z* number with zero:

a) if $\min\{b_z,c_z\}>0$ then $x>y$;

b) if $\max\{b_z,c_z\}<0$ then $x<y$;

c) if $\min\{b_z,c_z\}<0$ & $\max\{b_z,c_z\}<0$ & $|\min\{b_z,c_z\}|>\max\{b_z,c_z\}$ then $x<y$;

d) if $\min\{b_z,c_z\}<0$ & $\max\{b_z,c_z\}>0$ & $|\min\{b_z,c_z\}|<\max\{b_z,c_z\}$, then $x>y$.

An example of a possible result of the subtraction operation is shown in Fig. 1, 2, where on the left is a graphical description of the operands, and on the right is the result of the subtraction operation.

Experimental studies and performance evaluation of fuzzy cluster classifier using refined grouping procedure

Experimental studies and performance evaluation of fuzzy cluster classifier using refined grouping procedure were based on the analysis of CS status in two modes: normal and abnormal. Malicious software was used to simulate the abnormal state of CS.

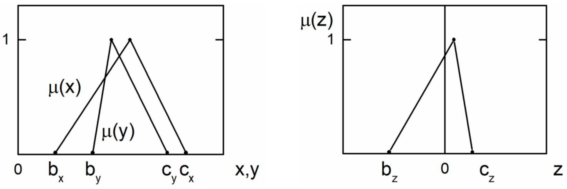

Fig. 1. The result of the calculation $x-y,|b_z|>c_z,x<y$ (rule c)

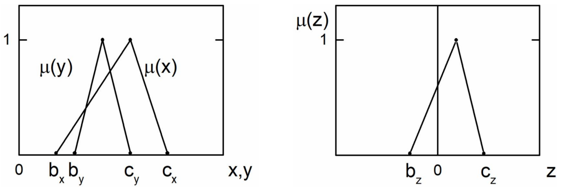

Fig. 2. The result of the calculation $x-y,|b_z|<c_z,x>y$ (rule d)

The developed software made it possible to obtain the main indicators of the functioning of the CS and to classify its state. The following performance indicators of the CS were used as source data: (CPU load, amount of memory used, network traffic volume, number of read/write operations to disk, invasion signatures, statistics on system events analysis, for example, the number of operations with the system registry or file system, the number of processes, etc.).

In order to evaluate the quality of classification, ROC analysis was used in this work. ROC analysis allows to evaluate the quality of diagnostic and prognostic methods and is widely used in binary classification problems [23, 24].

Fig. 3 shows a graph of the ROC curve for the process of classifying the state of a CS based on a fuzzy cluster classifier.

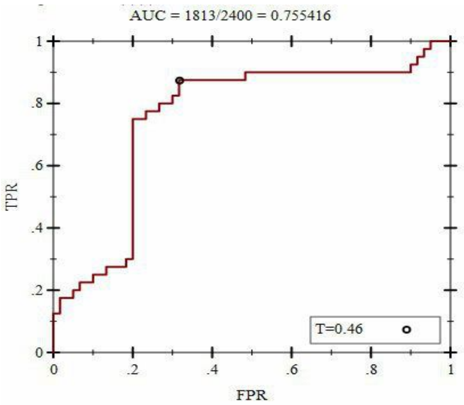

Fig. 3. The graph of the ROC curve for the process of classifying the state of a CS based on a fuzzy cluster classifier

TPR (the proportion of correctly classified events) is calculated as follows:

$$TPR=TP/(TP+FN), \qquad (3)$$

where TPR – the proportion of correctly classified events; FN – number of incorrectly classified negative events, TP – number of correctly classified positive events; FPR – the proportion of incorrectly classified events, is calculated as follows:

$$FPR = FP/(FP+TN)\$ \qquad (4)$$

TN – number of correctly classified negative events, FP – number of incorrectly classified positive events.

As can be seen from Fig. 3, the classification of the state of a CS based on a fuzzy cluster classifier is qualitative (the closer to 1 the area under the ROC curve is, the better quality of the classifier is).

In addition, the quality of the classification can be characterized by other indicators, namely: accuracy ACC (the proportion of correctly classified objects) and J statistics of Yuden (assessment of the probability of validity of the classification decision).

A comparative assessment of the identification quality of the CS state based on a standard and improved fuzzy cluster classifier is given in table 1.

Table 1 – **The assessment of the identification quality of the CS state based on a standard and improved fuzzy cluster classifier**

Method based on cluster analysis		
	Standard cluster classifier	Improved fuzzy cluster classifier
AUC	0,53	0,75
ACC	0,52	0,72
J-statistics	0,71	0,83

Conclusions

In this paper, a method for identifying the state of a CS with fuzzy initial data was theoretically grounded. This method involves the use of a fuzzy cluster classifier, which made it possible to perform identification of a computer system with a small sample of source data. The basis of the fuzzy cluster classifier is the developed simpler and more reliable procedure for comparing fuzzy numbers, based on comparing the fuzzy function of the difference in distance to the clustering center with zero. Software was developed that allowed to obtain the main indicators of the functioning of the CS and perform classification of the CS state. Indicators of the CS functioning were used as initial data: (CPU load, amount of memory used, network traffic volume, number of read/write operations to disk, invasion signatures, system events statistics, for example, the number of operations with the system registry or file system, the number of processes, etc.).

The classification quality assessment based on ROC analysis was carried out. It was found that the fuzzy cluster classifier is qualitative (AUC = 0.75) and allows to assess the state of the CS with a small sample of the initial data or if they are fuzzy. Studies of fluctuations in the proportion of correctly and incorrectly classified events depending on the value of the selected decision criterion (decision point, or cut-off point) were performed. This will allow to adjust the level of false positive and false negative classification of the CS state.

References

1. Shelukhin, O.I., Sakalema, D.Zh. and Filinova, A.S. (2013), *Intrusion detection in computer networks*, GIT, Moscow, 220 p.
2. Shkodyrev, P.V., Yagafarov, K.I., Bashtovenko, V.A. and Ilyin E.E. (2017), "A review of methods for detecting anomalies in data streams", *Proc. of the 2 Conf. on Software Engineering and Information Management*, St. Petersburg, Russia, Vol. 18, pp. 64–70.
3. Agrawal, S. (2015), "Survey on Anomaly Detection using Data Mining Techniques", *Proc. Computer. Science*, Vol. 60, pp. 708-713.
4. Chandola, V., Banerjee, A. and Kumar, V. (2012), "Anomaly detection for discrete sequences: A survey", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 823–839.
5. Barseghyan, A.A., Kupriyanov, M.S., Stepanenko, V.V. and Cold, I.I. (2007), *Data Analysis Technologies: Data Mining, Visual Mining, Text Mining, OLAP*, 2nd ed., Revised. And add, BHV-Petersburg, S-Pb., 384 p.
6. Semenov, S.G. and Gavrilenko, S.Yu. (2015), "Formation and study of heuristics in antivirus analyzers using the Mamdani algorithm", *Journal of Qafqaz university*, Azerbadhan, Mathematics and computer scienceVol. 3, No. 3, pp. 116-120.
7. Fisher, R. A. (1958), *Statistical methods for researchers*, Gosstatizdat, Moscow, 267 p.
8. Semenov, S., Gavrilenko, S. and Chelak, V. (2016), "Developing parametrical criterion for registering abnormal behavior in computer and telecommunication systems on basis of economic test", *Actual problems of economics*, Kyiv, Vol. 4(178), pp. 451-459.
9. Sutton, R.S. and Barto, A.G. (2020), *Reinforcement Learning*, 2-nd edition, DMK press, Moscow, 552 p.
10. Rokach, L. (2010), "Ensemble-based classifiers", *Artificial Intelligence Review*, Vol. 33, release 1–2.
11. (2020), *Methods of constructing decision trees in classification problems in Data Mining*, available at: https://ami.nstu.ru/~vms/lecture/data_mining/trees.htm
12. Iwan, Syarif, Ed, Zaluska1, Adam, Prugel-Bennett1 and Gary, Wills (2012), *Application of Bagging, Boosting and Stacking to Intrusion Detection*, Springer-Verlag Berlin Heidelberg. Perner (Ed.): MLDM, LNAI 7376, pp. 593–602.
13. Tarkhov, D.A. (2014), *Neural network models and algorithms*, Radio Engineering, Moscow, 352 p.
14. Barsky, A.B. (2004), *Neural networks: recognition, control, decision making*, Finance and statistics, Moscow, 176 p.
15. Rutkovskaya, D.S., Pilinsky, M.V. and Rutkovsky, L.P. (2004), *Neural networks, genetic algorithms and fuzzy systems*, Garyachaya liniya-Telecom, Moscow, 452 p.
16. Korchenko A.O. (2019), *Methods of identification of anomalous stations for systems of detection of intrusion*, Dis. doc those. 05.13.21 - Systems for information security, Kyiv, 405 p.
17. Lin, W-C., Ke, W-S. and Tsai, C-F (2015), "An intrusion detection system based on combining cluster centers and nearest neighbors", *Knowledge-Based Systems*, vol. 78, pp. 13-21.
18. Mandel, I.D. (1988), Methods of cluster analysis, Finance and Statistics, Moscow, 176 p.
19. Egorenko, M.V. and Bokhovko, A.G. (2016), "Cluster analysis as a means of grouping the studied variables", *Collection of St. Petersburg State University of Economics*, 2016, Issue 7, p. 57-69.
20. Kofman, A. (1982), *Introduction to the theory of fuzzy sets*, Radio and communications, Moscow, 486 p.
21. Semenov, S., Sira, O., Gavrylenko, S. and Kuchuk N. (2019), "Identification of the state of an object under conditions of fuzzy input data", *Eastern-European Journal of Enterprise Technologies*, vol. 1, no 4 (97), pp. 22-29, DOI: https://doi.org/10.15587/1729-4061.2019.1570.

22. Raskin, L.G. and O.V. Sira (2008), *Fuzzy math. Fundamentals of the theory. Applications*, Parus, Kharkiv, 352 p.
23. (2019), *Detector Performance Analysis Using ROC Curves,* available at: https://www.mathworks. com/help/phased/examples /detector-performance-analysis-using-roc-curves.html
24. Fawcett, T. (2006), “An Introduction to ROC Analysis”, *Pattern Recognition Letters*, 27 (8), pp. 861–874.

Received (Надійшла) 04.02.2020
Accepted for publication (Прийнята до друку) 20.05.2020

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Гавриленко Світлана Юріївна – доктор технічних наук, доцент, професор кафедри обчислювальної техніки та програмування, Національний технічний університет «Харківський політехнічний інститут»;
Svitlana Gavrylenko – Doctor of Technical Sciences, Associate Professor, Professor of the Department of " Computer Engineering and Programming ", National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gavrilenko08@gmail.com; ORCID ID: https://orcid.org/0000-0002-6919-0055

Челак Віктор Володимирович - асистент кафедри "Обчислювальна техніка та програмування", Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Viktor Chelak - Lecturer, Department of Computer Engineering and Programming, NTU “KhPI”, Kharkiv, Ukraine;
e-mail: victor.chelak@gmail.com; ORCID: https://orcid.org/0000-0001-8810-3394

Горносталь Олексій Андрійович – асистент кафедри "Обчислювальна техніка та програмування", Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Oleksii Hornostal - Lecturer, Department of Computer Engineering and Programming, NTU “KhPI”, Kharkiv, Ukraine;
e-mail: gornostalaa@gmail.com; ORCID: https://orcid.org/0000-0001-5820-9999

Велізар Ангелов Вассилев – кандидат технічних наук, доцент, кафедра Прецензійної техніки та приладобудування, Технічний університет Софія, Софія, Болгарія
Velizar Vassilev – PhD, Assistant Professor, Department of Precise Engineering and Measurement Instruments, Technical University of Sofia, Sofia, Bulgaria
e-mail: vassilev_v@tu-sofia.bg; ORCID: https://orcid.org/0000-0003-1563-2353

Розробка методу ідентифікації стану комп’ютерної системи з використанням нечіткого кластерного аналізу

С. Ю. Гавриленко, В. В. Челак, О. А. Горносталь, В. А. Вассилев

Анотація. Предметом статті є дослідження методів ідентифікації стану комп’ютерних системах. **Метою** статті є розробка методу ідентифікації аномального стану комп’ютерної системи на основі нечіткого кластерного аналізу. **Завдання**: проаналізувати методи ідентифікації стану комп’ютерних систем; провести дослідження з вибору вихідних даних; розробити метод ідентифікації стану комп’ютерної системи за умови малої вибірки або нечіткості вихідних даних; дослідити та обґрунтувати процедуру порівняння нечітких відстаней між центрами групування і об’єктами кластеризації; розробити програмне забезпечення та провести тестування. Використовуваними **методами** є: кластерний аналіз, апарат нечіткої логіки. Отримано такі **результати**: теоретично обґрунтовано та досліджено метод ідентифікації стану комп’ютерної системи за умови малої вибірки або нечіткості вихідних даних, який відрізняється використанням методу на основі нечіткого кластерного аналізу з уточненою процедурою групування. Для вирішення завдання кластеризації використана спеціальна процедура порівняння нечітких відстаней між центрами групування і об’єктами кластеризації. Розроблено програмне забезпечення та виконано тестування розробленого методу. Проведено оцінку якості класифікації на основі ROC-аналізу **Висновки**. Наукова новизна отриманих результатів полягає в наступному: проведено дослідження з вибору вихідних даних для аналізу; розроблено метод ідентифікації стану комп’ютерної системи на основі нечіткого кластерного аналізу з використанням спеціальної процедура порівняння нечітких відстаней між центрами групування і об’єктами кластеризації, що дозволило покращити якість класифікації до 22%.

Ключові слова: ідентифікація стану; комп’ютерна система; кластерний аналіз; нечіткі вихідні дані.

Разработка методу идентификации состояния компьютерной системы с использованием нечеткого кластерного анализа

С. Ю. Гавриленко, В. В. Челак, А. А. Горносталь, В. А. Вассилев

Аннотация. Предметом статьи является исследование методов идентификации состояния компьютерных системах. **Целью** статьи является разработка метода идентификации аномального состояния компьютерной системы на основе нечеткого кластерного анализа. **Задача**: проанализировать методы идентификации состояния компьютерных систем; провести исследования по выбору исходных данных; разработать метод идентификации состояния компьютерной системы при малой выборки или нечеткости исходных данных; исследовать и обосновать процедуру сравнения нечетких расстояний между центрами группировки и объектами кластеризации; разработать программное обеспечение и протестировать. Используемыми **методами** являются: кластерный анализ, аппарат нечеткой логики. Получены следующие **результаты**: теоретически обосновано и исследован метод идентификации состояния компьютерной системы при малой выборки или нечеткости исходных данных, который отличается использованием метода на основе нечеткого кластерного анализа уточненной процедурой группировки. Для решения задачи кластеризации использована специальная процедура сравнения нечетких расстояний между центрами группировки и объектами кластеризации. Разработано программное обеспечение и выполнено тестирование разработанного метода. Проведена оценка качества классификации на основе ROC-анализа. **Выводы**. Научная новизна полученных результатов заключается в следующем: проведено исследование по выбору исходных данных для анализа; разработан метод идентификации состояния компьютерной системы на основе нечеткого кластерного анализа с использованием специальной процедура сравнения нечетких расстояний между центрами группировки и объектами кластеризации, что позволило улучшить качество классификации до 22%.

Ключевые слова: идентификация состояния; компьютерная система; кластерный анализ; нечеткие выходные данные.