The analysis of information security problems solved by clustering methods

A R Aidinyan\textsuperscript{1}, O L Tsvetkova\textsuperscript{1}, A N Gerasimenko\textsuperscript{1} and O Ja Kravets\textsuperscript{2}

\textsuperscript{1} Don State Technical University, 1, Gagarina sq., Rostov-on-Don, 344003, Russian Federation
\textsuperscript{2} Voronezh State Technical University, 14, Moscow ave., Voronezh, 394026, Russian Federation

E-mail: andstyle@mail.ru

Abstract. The actual problem of information security is achieved by using a variety of methods and technologies. The level of information protection depends on the quality of the methods used. The article deals with information security problems, and provides a classification and analysis of clustering algorithms. The clustering methods considered during the classification process have certain advantages and disadvantages, the correct choice of which ensures their effective application in solving specific problems. Clustering methods k-means and k-medians in most cases give better results with the right choice of clustering criteria. These methods are simple, contain a small number of calculations, and avoid the problem of sensitivity to the initial choice of cluster centers. It also makes sense to use fuzzy clustering if cluster elements cannot be clearly assigned to one of the clusters.

1. Introduction and statement of the problem
The relevance of ensuring information security is beyond doubt. To ensure information security, various tasks are solved using various methods. The ability to use specific information and telecommunication services and applications at the enterprise depends on the quality of the applied information security technologies. In this regard, the task of solving the problem of developing, improving, and implementing information security technologies both in the field of cryptography, antivirus protection, and in the field of technical means is especially relevant.

In a number of information security tasks, it is possible to use clustering methods. The purpose of this article is to consider information security problems and select clustering algorithms suitable for a particular task.

2. Overview of clustering algorithms
Scientific research in the development of approaches and methods for performing cluster analysis is relevant and in demand in various scientific and applied fields. Cluster analysis is the task of splitting a given sample of objects into a number of subsets, called clusters. The similarity of objects can be expressed in different ways, and various measures of proximity are proposed to determine it [1].

When choosing a clustering method, it is necessary to take into account its features, since using different methods for the same set of source data will lead to different results.
Interesting approaches to the classification of clustering methods are given in [1, 2]. Figure 1 shows the classification of clustering methods from [1]. According to the method of data processing, all clustering methods are divided into two large groups: hierarchical and partitioning [1, 2].

With hierarchical clustering methods, smaller clusters are sequentially merged into larger ones, or large clusters are divided into smaller ones. Thus, hierarchical methods are divided into agglomerative and divisive methods [3, 4].

Partitioning methods are iterative and based on the choice of a representative of the cluster (for example, the arithmetic mean of the cluster objects) and as a result, these methods divide a set of objects into a given number of clusters $K$. At each step, objects that are closest to their representatives are added to the clusters and the representatives are re-selected. The main disadvantage of this group of methods is the sensitivity to noisy data, as well as the need for a priori setting of parameter $K$ [5].

An example of using the iterative clustering method is the k-means and k-median algorithms that build a given number of clusters located as far as possible from each other [6, 7]. The advantages of the k-means algorithm are simplicity and speed of implementation. The disadvantage of the k-means algorithm is that its speed decreases sharply with a large number of objects. The k-means algorithm is also sensitive to outliers that distort the average. To solve this problem, the k-median algorithm can be used.

Other popular algorithms that implement iterative clustering methods are PAM (k-means + k-medoids) [8], CLOPE, and LargeItem.

Density-based clustering methods define a cluster as a group located sufficiently heap objects. An example of the implementation of the density clustering method is the DBSCAN algorithm, which is based on the assumption that within each cluster a typical density of objects is observed, which is noticeably higher than the density outside the cluster.
Model-based methods provide a search and maximize the similarity between a given cluster model and the existing data set. The most popular method for representing models is the EM (Expectation-Maximization algorithm) [9–11].

The advantages of hierarchical methods in comparison with non-hierarchical methods are their visibility and the ability to get a detailed idea of the structure of the studied data. The disadvantages of hierarchical clustering methods include the limitation of the size of the dataset, the need to select the criterion of affinity and rigidity of derived classifications.

By the method of data analysis, clustering methods are divided into crisp and fuzzy. Crisp methods ensure the distribution of objects between clusters so that each object belongs to only one cluster. The use of fuzzy algorithms leads to the fact that each object belongs to each cluster with some probability [12]. The popular algorithm is Fuzzy c-means [13, 14].

Recently, genetic algorithms [15], neural networks [16, 17], fuzzy logic [18, 19], and other modern approaches have been used for clustering.

3. **The tasks of ensuring information security solved using clustering methods.**

The tasks of ensuring information security using clustering methods include (see figure 2):

- Clustering of information systems or their segments in order to divide them into nodes of the same type, which will allow assigning the same type of security policy to the elements of one cluster.
- Clustering the organization’s premises in order to divide them into groups and optimize costs while ensuring the required level of information security.
- Clustering information security incidents with the aim of analyzing and taking measures to optimally redistribute the costs of ensuring information security. Information security incidents can be classified by the criticality category of incidents, which can lead to significant negative
consequences for information assets or the organization’s reputation; to negative consequences; to minor negative consequences. Therefore, to optimize the costs of eliminating the costs of ensuring information security, it is necessary to classify the incidents that occurred [20, 21].

- Clustering files to identify threats [22], for example, when searching for computer viruses.
- Quantitative assessment of information security risks. The task of risk assessment is non-linear, complex, and has other restrictions on the use of traditional models for assessing information security risks [23-25]. Using clustering can simplify the task and, for example, determine the relationship between risk factors and the level of information security with the calculation of mutual information.

4. Conclusion

The clustering methods considered in the classification process have certain advantages and disadvantages that ensure their effective application in solving various problems. Studies have shown that the clustering methods k-means and k-medians in most cases give better results with the correctly choice of criteria for clustering. These methods are simple, involve little computation, and avoid the problem of sensitivity to the initial selection of cluster centers. It also makes sense to use fuzzy clustering if the cluster members cannot be clearly assigned to one of the clusters.

References

[1] Madhulatha T S 2012 An Overview on Clustering Methods IOSR Journal of Engineering 2(4) 719-25
[2] Aidinyan A R and Tsvetkova O L 2018 The cluster algorithms for solving problems with asymmetric proximity measures Numerical Analysis and Applications 11(2) 99-107
[3] Sneath P H A and Sokal R R 1973 Numerical taxonomy: The Principles and Practices of Numerical Classification (London: UK Freeman) p 573
[4] King B J 1967 Am. Stat. Assoc. 69 86-101
[5] Ester M, Kriegel H-P, Sander J and Xu X 1998 Clustering for mining in large spatial databases KI Künstliche Intell 12 18-24
[6] Steinhaus H 1957 Sur la division des corps matériels en parties Bulletin L’Académie Polonaise des Sciences 4 801-4
[7] Huang Z 1998 Extensions to the k-means algorithm for clustering large data sets with categorical values Data Mining and Knowledge Discovery 2 283-304
[8] Kaufman L and Rousseuw P J 2005 Finding Groups in Data: An Introduction to Cluster Analysis (John Wiley & Sons Inc.) p 342
[9] Hogg R, McKean J and Craig A 2005 Introduction to mathematical statistics (Upper Saddle River, NJ: Pearson/Prentice Hall) p 64
[10] MacKay D J C 2003 Information Theory, Inference, and Learning Algorithms (Cambridge: Cambridge University Press) p 640
[11] Dempster A P, Laird N and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm Journal of the Royal Statistical Society B 39(1) 1-38
[12] Yin B and Songhua H 2008 Fuzzy K-prototypes clustering based on particle swarm optimization Comput Eng Des 29(11) 2883-5
[13] Ruspini E H 1969 A new approach to clustering Information and Control 15 22-32
[14] Bezdek J C 1981 Pattern recognition with fuzzy (New York: Plenum Press) p 272
[15] Aidinyan A R and Tsvetkova O L 2011 Genetic algorithms of work distribution Vestnik of Don State Technical University 11(5) 723-9
[16] Krasnosnichkov A 2010 Complex data clustering with single-layer dynamically linked spiking neural network Vestnik of Don State Technical University 10(3) 318-24
[17] Anikin V I and Karmanova A A 2015 Clustering and classification of multidimensional data by the Kohonen’s cellular neural network Informacionnye technologii 6 470-7
[18] Xu L and Zhang F 2006 Fuzzy clustering algorithm based on PSO Comput Eng Des 27(21) 4128-
[19] Zhang D, Liu X and Guan Z 2006 A dynamic clustering algorithm based on PSO and its application in fuzzy identification *International Conference on Intelligent Information Hiding and Multimedia* 232-5

[20] Zaytsev A S and Malyuk A A 2016 Development of information security insider threat classification using incident clustering *Bezopasnost’ Informatsionnykh Tekhnologiy* 23(3) 20-9

[21] Kurinnykh D Yu, Aydinyan A R and Tsvetkova O L 2018 The approach of clustering threats to information security of enterprises *Inženernyj vestnik Dona* 1(91) http://ivdon.ru/uploads/article/pdf/IVD_188_Kurinnykh_Ajdinyan_TSvetkova.pdf_b9f8256daa.pdf

[22] Kiselev Yu A 2012 Prospects for the use of genre classification of web documents in search engines *Inženernyj vestnik Dona* (Rus) 4(2) http://www.ivdon.ru/ru/magazine/archive/n4p2y2012/1425/

[23] Gao G, Li X, Zhang B and Xiao W 2011 Information security risk assessment based on information measure and fuzzy clustering *Journal of Software* 6 2159-66

[24] Tsaregorodtsev A V, Kravets O Ja, Choporov O N and Zelenina A N 2018 Information Security Risk Estimation for Cloud Infrastructure *International Journal on Information Technologies and Security* 4(10) 67-76

[25] Romansky R 2019 A Survey of Informatization and Privacy in the Digital Age and Basic Principles of the New Regulation *International Journal on Information Technologies and Security* 1(11) 95-106