УДК 681.3.093:044.3

DENIS OLEGOVICH ZUEV
Independent Consultant Lead Arcitect, Network and Cloud USA, Colorado
ARTEMII VASILYEVICH KROPACHEV
Bell Integrator USA Automation Solution Department Manager USA, Colorado
ALEKSEY YEVGENYEVICH USOV
Technical Architect Russian Govt Insurance Russia, Moscow
DMITRII NIKOLAEVICH MOSTOVSHCHIKOV
Bell Integrator Russia, Moscow System Installation Solutions Department Manager

ANALYSIS OF INFORMATION SYSTEMS SECURITY ALGORITHMS BASED ON CLUSTERING METHODS

Possible ways of application of clustering algorithms in the security of network and information systems area were considered. Information systems network transparency support, hardware resources release development and clustering security algorithms procedures were discussed. It was shown that cyber-defense systems are often slow down network service performance so it's necessary to develop algorithms which would not create extra overheads. Generally intrusion detection model was analyzed as system which includes functional blocks of misuse intrusion detection and anomaly intrusion detection. Most common and convenient similarity metric systems to use at modern network clustering algorithms were discussed. It was analyzed how to use in the area of clustering algorithms development Minkowski Distance, City Block Distance and Mahalanobis Distance metrics'. Hierarchal, partitional, density-based and grid-based algorithms of data clustering were proposed and analyzed. It was shown that most effective hierarchal clustering algorithms are BIRCH, CURE and ROCK, most effective partitional clustering algorithms are K-Means; CLARA and CLARANS as version of CLARA, most effective density-based clustering algorithms are DBSCAN, DBCLASD, GDBSCAN DENCLUE and OPTICS, while most effective grid-based clustering algorithms are STING, CLIQUE, GRIDCLUS, Wave Cluster and OptiGrid. It was mentioned that main benefits of hierarchical clustering methods are flexibility in adaptation of any metrics systems type of attribute and possibility to work with undefined set of data which is very useful in order to work with nowadays scalable network services. It was shown that range of application of clustering algorithms is based on amount of data to be analyzed and hardware resources of platform. An algorithms based on each clustering method that is proved to be most effective was demonstrated. A mathematical model for determining the efficiency of security strategy based on clustering algorithms was build and discussed.

Keywords: *network security, information system, clustering method, hierarchal algorithm, partitional algorithm, density-based algorithm, grid-based algorithm.*

ДЕНИС ОЛЕГОВИЧ ЗУЕВ
независимый консультант, ведущий архитектор сетей и облачных вычислений США, Колорадо,
АРТЕМИЙ ВАСИЛЬЕВИЧ КРОПАЧЕВ
руководитель департемента решений автоматизации Bell Integrator USA США, Колорадо,
АЛЕКСЕЙ ЕВГЕНЬЕВИЧ УСОВ
технический архітектор ПАО СК "Росгосстрах" Россия, Москва,
ДМИТРИЙ НИКОЛАЕВИЧ МОСТОВЩИКОВ
руководитель отдела решений системной инсталляции Bell Integrator Россия, Москва

АНАЛИЗ АЛГОРИТМОВ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ, ОСНОВАННЫХ НА МЕТОДАХ КЛАСТЕРИЗАЦИИ

Рассмотрены возможные способы применения алгоритмов кластеризации в области безопасности сетей и информационных систем. Изучены процедуры обеспечения прозрачности сетевого трафика, освобождения аппаратных мощностей и разработки кластеризационных защитных алгоритмов. Было показано, что системы кибер-защиты зачастую снижают производительность сетевых сервисов, поэтому необходимо разработать алгоритмы, которые будут оказывать минимальное влияние на распределение аппаратных ресурсов. Рассмотрена базовая модель обнаружения несанкционированного внедрения как система, которая включает в себя функциональные элементы определения внутренней атаки и атаки произведенной извне. Изучены наиболее распространенные и удобные метрические системы определения сходства объектов, которые могут быть использованы в современных алгоритмах кластеризации сетевых ресурсов. Проанализировано, как использовать в области разработки алгоритмов кластеризации метрику Минковского, Городских Кварталов и Махаланобиса. Предложены и проанализированы иерархические, разделяющие, плотностные и грид-алгоритмы кластеризации данных. Было показано, что наиболее эффективными иерархическими алгоритмами кластеризации являются BIRCH, CURE и ROCK, наиболее эффективными разделяющими алгоритмами кластеризации являются K-Means; CLARA и CLARANS как версия CLARA, наиболее эффективными плотностными алгоритмами кластеризации являются DBSCAN, DBCLASD, GDBSCAN DENCLUE и OPTICS, в то время как наиболее эффективными грид-алгоритмами кластеризации являются STING, CLIQUE, GRIDCLUS, Wave Cluster и OptiGrid. Отмечено, что основными преимуществами иерархических методов кластеризации являются гибкость в адаптации, а также возможность работы с неопределенным набором данных, что очень важно для работы с современными масштабируемыми сетевыми сервисами. Предложены методы кластеризации больших наборов данных, отмечено, что область применения алгоритмов кластеризации зависит количестве анализируемых объектов и аппаратных ресурсов платформы. Была построена и проанализирована математическая модель для определения эффективности построения алгоритмов обеспечения кибер-безопасности сетевых сервисов.

Ключевые слова: *сетевая безопасность, метод кластеризации, алгоритм иерархической кластеризации, алгоритм частичной кластеризации, алгоритм плотностной кластеризации, алгоритм грид-кластеризации.*

Introduction. Clustering is an unsupervised summarization technique of gathering similar objects into groups (clusters). The similarity among objects has to be determined through parameter called metrics. Information systems' (IS) developers use clustering methods to optimize hardware and software environment by components sorting. Application of clustering algorithms for support of IS security includes following procedures [1-4]:

- IS network transparency support;
- IS hardware resources release;
- development of clustering security algorithms.

As it shown at Fig. 1 support of the network processes transparency (i.e. classification of all allowed processes) helps to clarify IS security policies, sorting of hardware environment components optimizes IS infrastructure allowing to release some resources for security platform and, finally, clustering methods could be directly used in cyber-defense algorithms.

Generally intrusion detection system (IDS) model includes two functional blocks:

- misuse intrusion detection (MID);
- anomaly intrusion detection (AID).

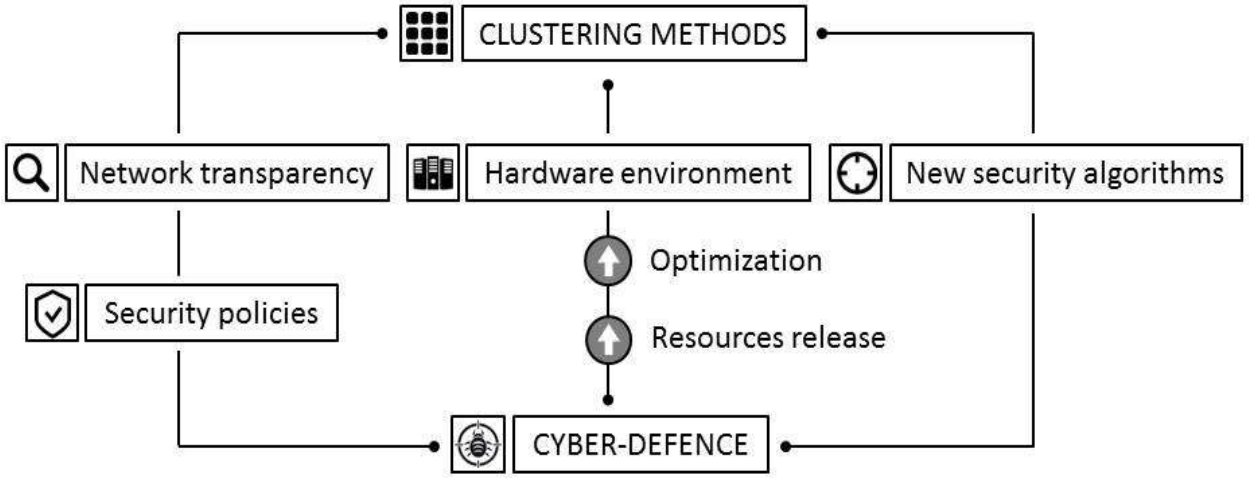

Fig. 1. Application of clustering algorithms for support of IS securit

Effective detection of MID-class treats is all to accurate security policies development while detection of AID caused by global network activity is often a nontrivial task. Major part of cyber-treats could be detected by active monitoring tools. Clustering algorithms are usually to be used at the stage of passive monitoring, specifically during data mining process.

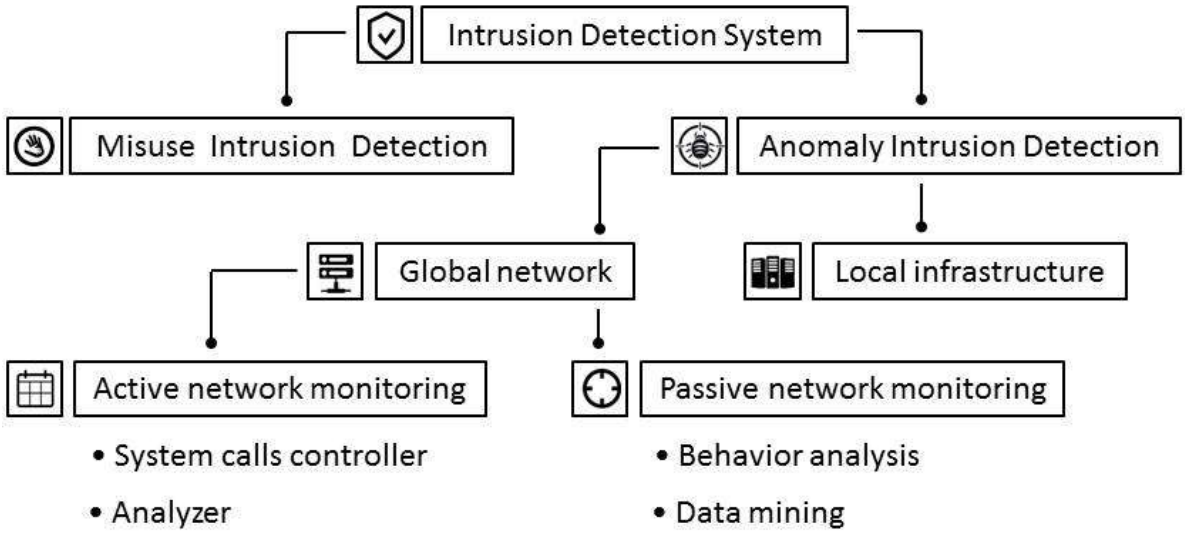

Fig. 2. Basic components of intrusion detection system

It should be noticed that development of indirect as well as direct methods of security system which is based on clustering algorithms requires deep understanding of clustering methodology.

1. Clustering methodology. As it was noticed clustering algorithm are based on metrics which determined similarity among objects. Let's analyze similarity among objects x_{il} and x_{jl} which could be characterized by d set of parameters ($l \in [1; d]$):

$$\begin{cases} x_{il} = (x_{i1} \dots x_{id}) \\ x_{jl} = (x_{j1} \dots x_{jd}) \end{cases}. \quad (1)$$

Similarity among x_{il} and x_{jl} could be defined by distance value D_{ij} which is obtained by mathematical equations of chosen metrics. In the area IT researchers usually adopt following metrics' systems (Fig. 3):

- Minkowski Distance;
- Manhattan or City Block Distance;
- Mahalanobis Distance.

It should be mentioned that for Minkowski Distance equation researchers often set value $n = 2$ which corresponds to Euclidean Distance (Fig. 3).

As for clustering of categorical data similarity measure will be different. For two categorical data points x_{il} and x_{jl} with l attributes it will be calculated as follows:

$$D(x_{il}, y_{il}) = \sum_{i=1}^{d} \delta(x_{il}, y_{il}), \quad (2)$$

where $\delta(x_{il}, y_{il})$ could be obtained as:

$$\delta(x_{il}, y_{il}) = \begin{cases} 1 \ \ if \ \ x_{il} = y_{il} \\ 0 \ \ if \ \ x_{il} \neq y_{il} \end{cases}. \quad (3)$$

Of course there are more complicated similarity measures; however, explained metric systems are most common and convenient ones to use at modern network clustering algorithms.

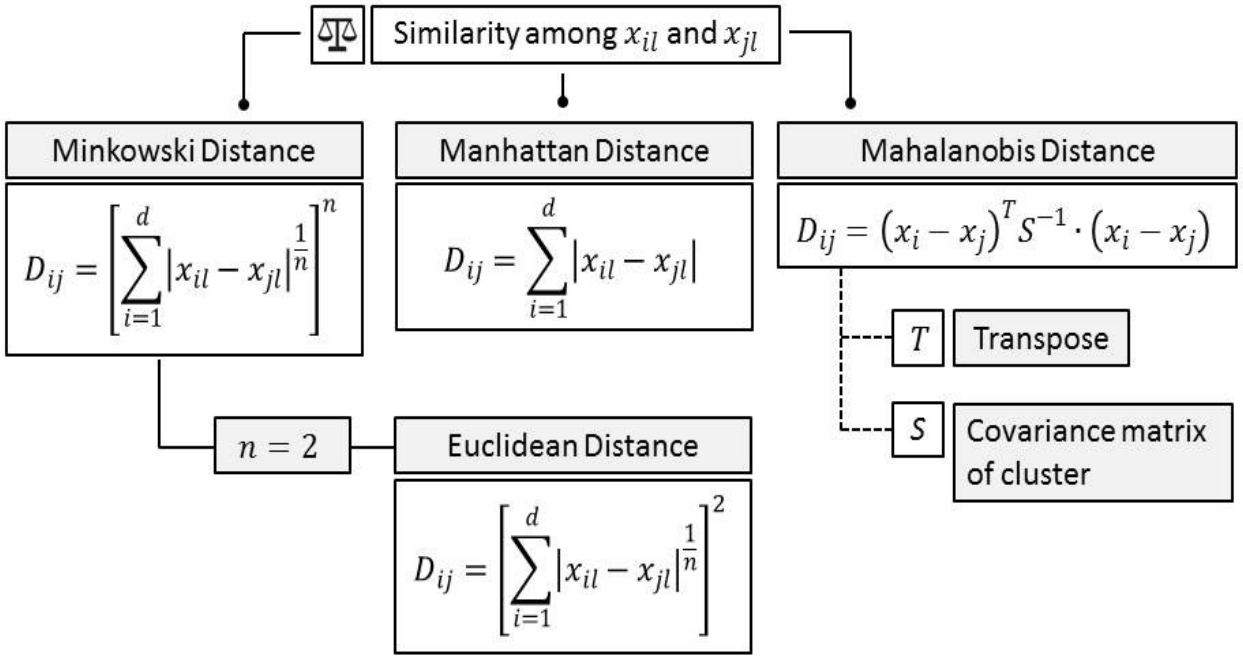

Fig. 3. Determination of similarity among objects by the metrics systems

Another important stage of clustering algorithm design is selection of clustering method. There are five groups of method to be used for clustering of network data [5-7]:

- hierarchical clustering;

- partitioning clustering;
- density-based clustering;
- grid-based clustering;

Selection of the method allows to determine type of data to be used in clusters, clusters shape, number of clusters and maximal size of data set to be clustered.

2. Clustering algorithms. Hierarchical clustering is a method that implies building of tree-structure of clusters where each cluster is represented as a node. This method is based on similarity measure (distance) and sort nearby objects (or nearby clusters, group of clusters, etc.) into a group. Hierarchal clustering algorithms form two major groups:

- bottom up (divisive);
- top down (agglomerative).

Bottom up hierarchical clustering starts from the cluster of entire data set (called "root") which should be divided into partitions up to the minimal cluster (called "leaves"). Top down hierarchical clustering, in other hands, starts from the elementary data elements (leaves). Set of these elements which should be formed into the clusters and the groups of clusters until entire data set cluster (root) will be formed (Fig. 4).

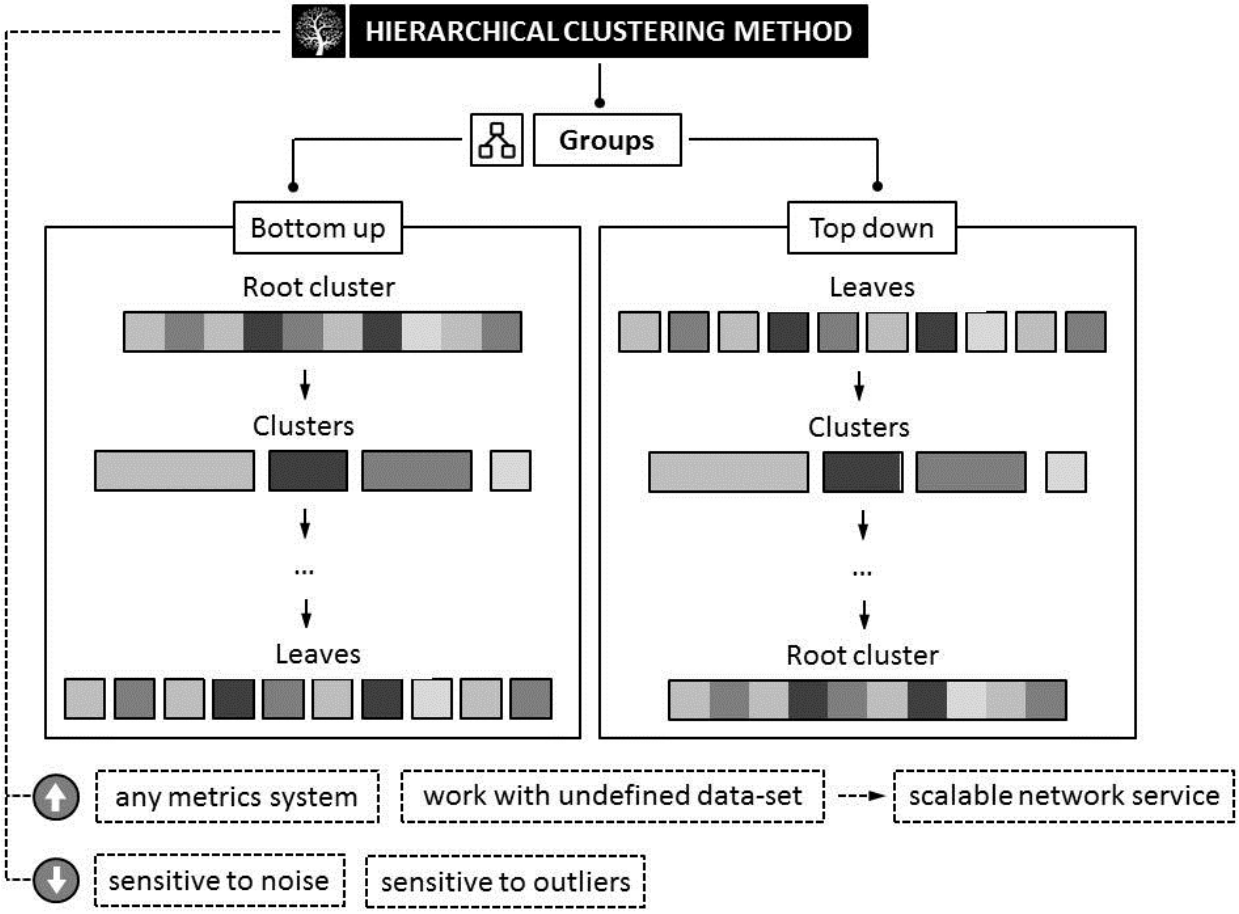

Fig. 4. Diagram of hierarchical clustering method

Main benefits of hierarchical clustering method are flexibility in adaptation of any metrics systems type of attribute and possibility to work with undefined set of data which is typical for nowadays scalable network services. In other hand, this method of clustering is very sensitive to noise and outliers, often incapable to correct misclassification and has no clear termination criterion. Thus modern hierarchal clustering algorithms should be able to cover mentioned disadvantages. Algorithms from this group that was proved to be most effective (Fig. 5) are:

- BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies);
- CURE (Clustering Using REpresentatives);
- ROCK (RObust Clustering using linKs)

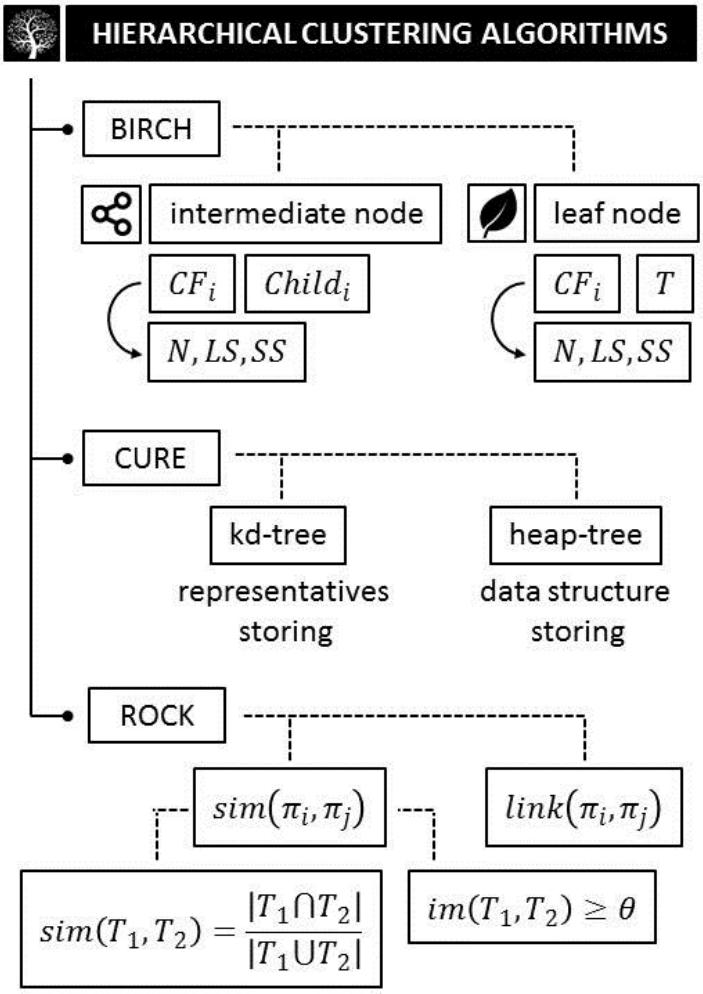

Fig. 5. Diagram of hierarchical clustering algorithms

BIRCH is an unsupervised data mining algorithm used to large data-sets. Major benefit of BIRCH-algorithm is incrementally and dynamically clustering of incoming data. It allows to get high quality of clustering and hardware resources consuming due to time constraints. BIRCH makes clustering decisions without scanning the whole data-set; it handles sparse area as outliers and removes them. The algorithm has a clustering feature tree structure (CFT-structure) with leaves and intermediate nodes characterized by entries. The number of entries must be constrained by parameters B (maximum number of intermediate node entries) and L (maximum number of leaf node entries). Entry of intermediate node could be defined as

$$\begin{cases} E_i = [CF_i, Ch_i] \\ CF_i \in (N, LS, SS) \end{cases}, \tag{4}$$

where CF_i is cluster data-points a summary information, Ch_i is a pointer to i^{th} child node, N is the number of data-points in a cluster, LS is the linear sum of the N data-points in a cluster and SS is the square sum of the N data-points in a cluster. Obviously entry of leaf node can be defined in a similar way:

$$\begin{cases} E_i = [CF_i, T] \\ T \in [1; T_{max}] \end{cases}, \tag{5}$$

where T is number of the leaves (inversely proportional quantity to height of the tree).

Algorithm CURE is very stable to outliers while it works with arbitrary-shape (e.g. non-convex) clusters. It's build to work with large data-sets with random sampling and partitioning procedures. First samples of data-sets in CURE-algorithm are to be chosen randomly and have to be partitioned to K equal partitions. To simplify the procedures first partitions could be clustered by BIRCH-algorithms. The algorithm ends with assigning label to each data points corresponding to distance from representatives. Thus, CURE-algorithm includes two data structures:

- kd-tree for representatives storing;
- heap-tree clusters for data structure storing.

The time complexity of the algorithms depends both on the number of sampling data and the number of partitions.

ROCK is an agglomerative hierarchical clustering algorithm that clusters data points similar to CURE-algorithm; however ROCK uses links instead of distance measure. ROCK also handles outliers and proved to be highly effective with very large data sets processing.

There are two similarities metrics to be used in ROCK to enable accurate merging and clustering of data points.

- $sim(\pi_i, \pi_j)$ as similarity measure to consider neighbors of a point;
- $link(\pi_i, \pi_j)$ to define the number of common neighbors between π_i and π_j.

$sim(\pi_i, \pi_j)$ could be defined as follow:

$$\begin{cases} sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \\ sim(T_1, T_2) \geq \theta \\ 0 \leq \theta \leq 1 \end{cases}. \quad (6)$$

where $|T_1|$ and $|T_2|$ are numbers of items in the transactions T_1 and T_2; θ is user-specified parameter based on closeness which shows if there are any similarity between transactions ($\theta = 0$ if any pair of transactions can be neighbors and $\theta = 1$ if only identical transactions can be considered as neighbors). The number of links between points indicates the probability whether data points are presented in a same cluster.

Partitioning clustering method includes partitioning data-set into k groups with n objects. Objective function of this method is minimizing square error function which is not effective for large data-sets due to high complexity of computation. Partitioning algorithms that are most widely used in IS are follows:

- K-Means;
- CLARA(Clustering LARge Applications);
- CLARANS(Clustering Large Applications based upon Randomized Search).

K-Means algorithm divides data objects into k partitions in order to assign objects to the nearest cluster centers. Value of k, cluster initialization and metric could be defined by researcher. K-means main goal is minimizing the within-cluster sum of square. It should be classified as a greedy algorithm with time complexity $O(N, T, k)$ where N is the number of objects and T is the number of algorithm iterations. K-means algorithm is scalable and could be used for clustering of large data-sets, but numbers of clusters in this case are needed to be defined. Practical issues also demonstrate K-means limitations at processing of outliers. CLARA algorithm could be considered as next level of PAM-algorithm (Partitioning Around Medoid) designed to solve problem of work with large data-sets. It has same as PAM time complexity of $O(k(n-k)^2)$. CLARANS is an advanced version of CLARA as for efficiency and scalability.

In Density-based algorithms, clusters are created based on highly dens areas over the remainder areas and the sparse area are classified as noise or border area. In this way, they can deal with outliers and non-convex shape clusters. Some of the mostly used density-based algorithms for large data-sets are follows:

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise);
- DBCLASD (Distribution Based Clustering of Large Spatial Databases);
- GDBSCAN (Generalized Density Based Spatial Clustering of Applications with Noise);
- DENCLUE (DENsity-based CLUstEring);
- OPTICS (Ordering Points To Identify the Clustering Structure).

DBSCAN algorithm clusters data-set in order reachability parameter. Data point directly density-reachable from another point if this is not far away than a given by user distance, so minimum number of points would be critical factors to generate a cluster. DBCLASD uses a uniform distribution of data points in a cluster. In this case nearest neighbor distance is a key parameter of cluster forming. This algorithm does not require loading the whole data-set while building a cluster so it's practical to use in network services. GDBSCAN is an improved version of DBSCAN. This algorithm changes the definition of neighborhood by binary predicate which is symmetric and reflexive, so GDBSCAN can define a neighborhood like intersect predicate. DENCLUE is based on h influence function. Sum of influence functions are calculated to obtain local maxima of the overall density function for defining of clusters, so algorithm is stable against noise and outliers. OPTICS is an algorithm for finding density-based clusters similar to DBSCAN. OPTICS-algorithm was designed to DBSCAN's main problem of detection of meaningful clusters in variable density data-set.

Grid-based clustering method is based on generation of finite number of cells by dividing data space and forming of grid structure. Major benefit of this algorithms is high velocity since they are dependent not on the number of data objects but on the number of cells. Basic grid-based clustering includes construction of grid structure

of non-overlapping cells, computing cells' density, sorting of the cells by the density value and identifying cluster centers. Mostly effective grid-based algorithms are follows:

- STING (Statistical Information Grid-based clustering);
- CLIQUE (Clustering in QUEst);
- GRIDCLUS;
- Wave Cluster;
- OptiGrid.

STING-algorithm clusters data-set into rectangular cells and forms hierarchical tree. It summaries data and store statistical information in each cell. Time complexity relies on number of grid cells at the lowest level. WAVECLUSTER transforms data-set into a frequency domain to find a dense area in it, so clusters with different resolutions and scales could be obtained. The computational complexity of WAVECLUSTER depends on number of objects in the data-set. GRIDCLUS neighbor search algorithm includes insertion of points into the grid structure, computing of density indices, sorting the blocks up to their density, identification of cluster centers, and traversal of neighbor clusters. OptiGrid algorithm is applied for high dimensional data-sets. It divides the whole data-set recursively into different subsets to find optimal grid partitioning. CLIQUE identifies subsets of a high dimensional data-set to achieve better clustering than original set. To find dense regions in a subset, each dimension is divided into equal intervals and area of high density should be found when the number of data points in this area exceeds threshold value defined by user.

3. Cyber-defense of information system based on clustering algorithms. To construct the model of IS cyber-defense platform; let's consider P as a set that includes all possible variants of the program code that can work in IS, $c \in C$ is the code to be analyzed, and $v \in C$ is a malware code. The definition of the cyber-attack can be defined as follows:

$$f(v_i) = f(c_i), \tag{7}$$

where $f(v_i)$ та $f(c_i)$ are functions of program code behavior defined by security policies, active monitoring and data mining (Fig. 6).

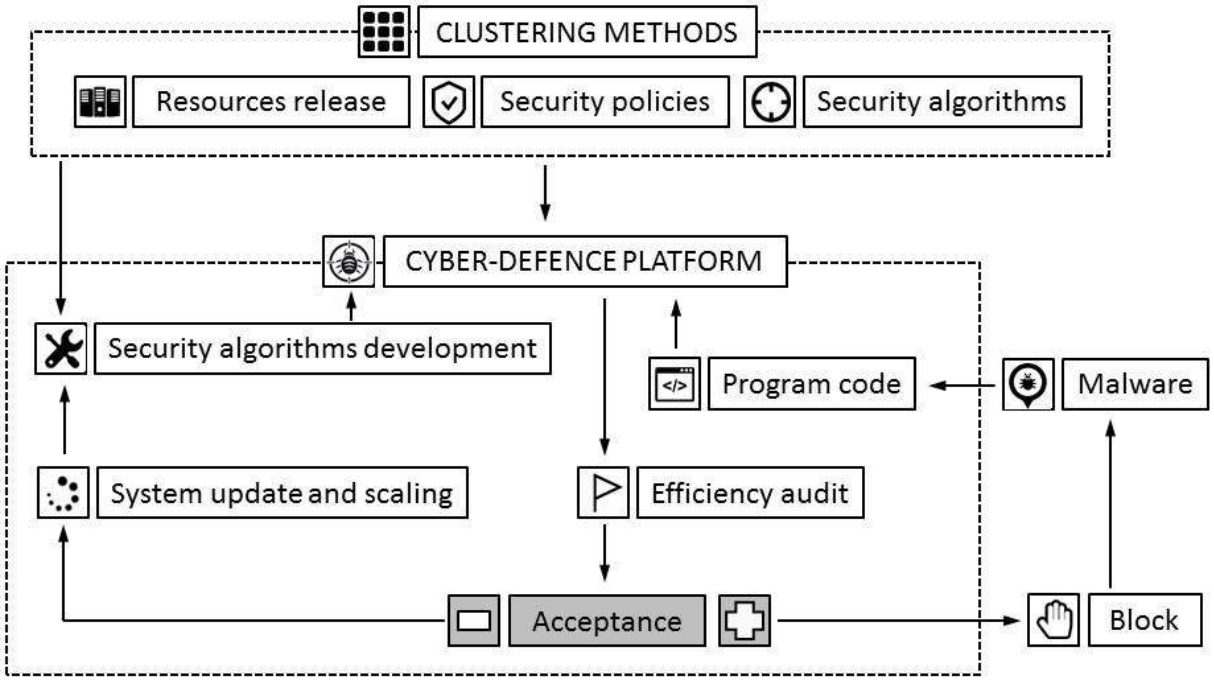

Fig. 6. Model of information system cyber-defense platform based on clustering algorithms

To define time interval $\Delta t = \tau_1 - \tau_0$ cyber-attack probability we should use integral function $T(v, c_i, a, t, s)$ which will define behavior of virus code v at IS s under constraints of additional factors a:

$$T(v, c_i, e, t, s) = \log \frac{\int_{\tau_0}^{\tau_1} f(v,a,t,s)dt}{\int_{\tau_0}^{\tau_1} c(p_i,a,t,s)dt}. \quad (8)$$

Thus, equation

$$T(v, c_i, e, t, s) \to 0 \quad (9)$$

could serve as an effective criterion of cyber-attack during Δt interval.

Proposed mathematical model provides the methodology for development of cyber-defense platform based on clustering algorithms.

Conclusions. There were considered methodology of application of clustering methods in the algorithms of cyber-security of network infrastructure and information systems. It was shown that cyber-defense methods should not significantly slow down network service performance and it's necessary to develop algorithms which do not create extra overheads. Hierarchal, partitional, density-based and grid-based algorithms of data clustering were proposed and analyzed. It was shown that range of application of clustering algorithms is based on amount of data to be analyzed and hardware resources of platform. A mathematical model for determining the efficiency of security strategy based on clustering methods was build and discussed.

References

1. Y. Mo, T. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopol, "Cyber-physical security of a smart grid infrastructure," Proceedings of the IEEE, vol. 100, no. 1, pp. 195-209, January 2012.
2. Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, "On false data-injection attacks against power system state estimation: Modeling and countermeasures," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 3, pp. 717 - 729, March 2014.
3. L. L. an M. Esmalifalak, Q. Ding, V. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," IEEE Transactions on Smart Grid, vol. 5, pp. 612 - 621, March 2014.
4. Peters, Wende. "Integrated Adaptive Cyber Defense." Proceedings of the 2015 Workshop on Automated Decision Making for Active Cyber Defense - SafeConfig 15, 2015, doi:10.1145/2809826.2809827.
5. Dumitraş, Tudor. "Understanding the Vulnerability Lifecycle for Risk Assessment and Defense Against Sophisticated Cyber Attacks." Advances in Information Security Cyber Warfare, 2015, pp. 265–285., doi:10.1007/978-3-319-14039-1_13.
6. Xu, Guandong, et al. Applied data mining. CRC Press, 2013.
7. Gan, Guojun, et al. Data clustering: theory, algorithms, and applications. SIAM, Society for Industrial and Applied Mathematics, 2007.

Рецензія/Peer review : 28.1.2018 р. Надрукована/Printed :9.4.2018 р.
Рецензент :