ᲡᲐᲥᲐᲠᲗᲕᲔᲚᲝᲡ ᲛᲔᲪᲜᲘᲔᲠᲔᲑᲐᲗᲐ ᲔᲠᲝᲕᲜᲣᲚᲘ ᲐᲙᲐᲓᲔᲛᲘᲘᲡ ᲛᲝᲐᲛᲑᲔ, ტ. 15, №4, 2021
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 15, no. 4, 2021

Informatics

Network Anomaly Detection in Corporate Security

Gulnara Janelidze*, Badri Meparishvili*, Luka Shonia*

* *Faculty of Informatics and Management Systems, Georgian Technical University, Tbilisi, Georgia*

(Presented by Academy Member Gocha Chogovadze)

The Corporate Security quality depends on Network Security, which is one of the vital assets in most corporations and should be well protected through effective information security practices. We focus on some corporate security challenges, especially hierarchical clustering problem as a Network Anomaly Detection problem and opportunity. The K-means algorithm combined with genetic algorithm for clustering have high performance quality of clustering with minimum time of process convergence. We consider entropy as an evaluation metric for our clustering. As data from devices, applications, and users are very large, Big-data technologies give opportunities of storage and analytical processing of large structured and unstructured data sets. Our approach is based on a rough idea that main data source of network supervising for the purpose of anomaly detection is transactional logs of whole corporation. Data file stream of events constantly fixed in transactional logs is unstructured big data, which require parallel storage and analytical processing using Hadoop Mapreduce framework. As a final output we get a list identified of network anomalies for further decisions. *© 2021 Bull. Georg. Natl. Acad. Sci.*

Clustering, K-means, genetic algorithms, entropy, Mapreduce framework

Many business corporations are facing a big challenge of transforming themselves into digital corporations while dealing with legacy IT systems and real external circumstances. Modern corporate governance, as a multilevel hierarchical system, means the effectiveness of an organization and management of employees within an organization. As information is a fundamental organizational notion, its security, together with other physical and technological means, must be integrated into the organization’s overall management plan. In the given context, information network security needs to be implemented and managed within the organization to ensure that the information is kept safe and secure.

The Corporate Information Security Policy should contain the following types of security: Physical Security, Network Security and Operational Security. Network Security has a vital importance for most corporations. With the rapid expansion of Internet technology and the accompanying, increasing number of network attacks, the network intrusion detection has become an important research area. The notion anomaly detection indicates to the problem of discovery nonconformity of patterns in network traffic data [1].

Materials and Methods

There are number of methods introduced for identification of network anomalies by monitoring and analyzing network traffic. Most methods identify anomalies by finding deviations from a normal traffic model. Existing anomaly detection methods are categorized into some categories [2].

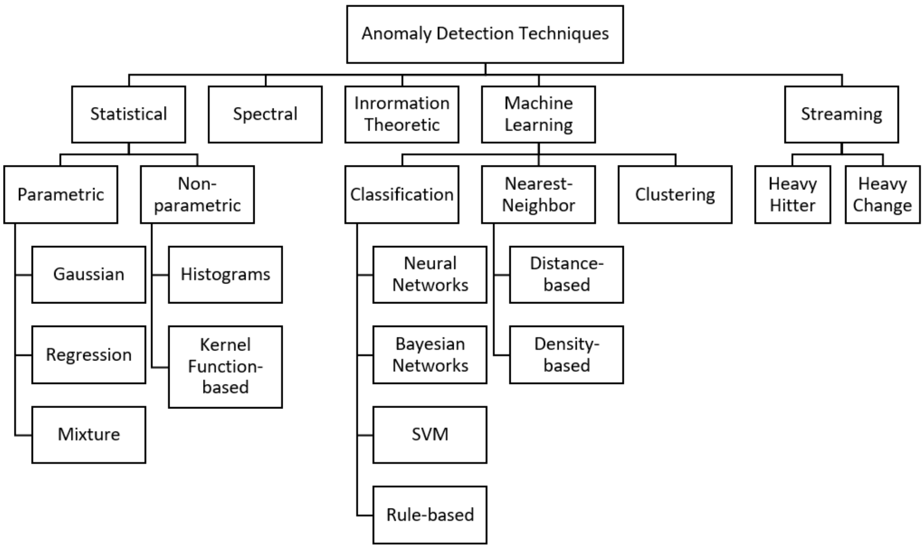

Fig. 1. Classification of anomaly detection methods.

In this paper, we discuss some corporate security challenges and how Machine Learning Methods, especially Clustering problem should be considered a Network Anomaly Detection problem and opportunity. The essence of clustering or cluster analysis consists in division of data objects set into groups that are formally called clusters. The objects within a cluster are more similar to each other than those in other clusters. Clustering is a type of unsupervised learning. Clustering methods group data into clusters-based on a similarity measure or distance computation. Clustering has the ability to detect anomalies without requiring explicit explanations of classes or types of anomalies. The k-means algorithm [3] is one of the most well-known centroid algorithms. It partitions the dataset into k subsets such that all points in a given subset are close to the same center. It randomly selects k of the instances as the cluster centers and all remaining instances are assigned to their nearest cluster center. k-means then computes the new cluster centers by taking the mean of all data points belonging to the same cluster.

Many research has been concentrated out on K-means combined with genetic algorithm for clustering. The problem on the large scale was divided into multiple problems. Almost K-means algorithm with GA have high performance quality of clustering with minimum time of process convergence [4].

We also discuss computing algorithms which use ideas from probability along with statistical computation to acquire patterns within the data. Soft computing methods which include approaches such as genetic algorithms, neural networks, fuzzy sets and rough sets, are suitable for network anomaly detection because often one cannot find exact solutions. So computing method such as genetic algorithms can be an acceptable method for network anomaly detection.

Genetic algorithms (GAs) are based on principles of evolution and natural selection. The Data structure is converted into so-called a chromosomes are evolved through many generations using operations such as selection, recombination and mutation. In the network anomaly detection problem, a chromosome for an individual contains genes corresponding to attributes such as services, ages, logged in or not and the number of superuser attempts. In computer network security applications, evolutionary computing is used mainly for anomaly finding as solutions to optimization problems.

A novel clustering technique is presented that combines both K-means and genetic algorithm (GA) together aims to gain better quality clusters without any need for user inputs like number of clusters k [5]. This hybrid approach is successful to finding the right number of clusters and identifying the right genes via a novel initial population determination approach. With the help of their novel fitness function and rearrangement operation of gene, it results in advanced quality cluster centers. The combination of K-means and readjustment of gene brings better outcome than K-means by itself.

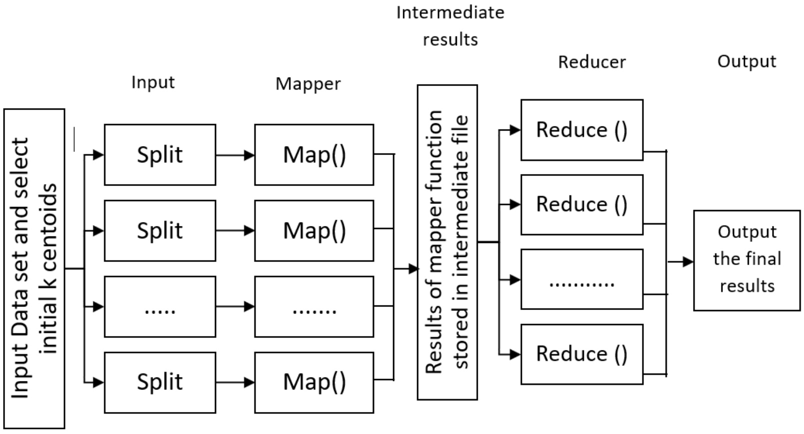

Fig. 2. The principal workflow of Log file processing on Mapreduce.

Proposed Aproach

The main point of Corporate Security is to transform the huge data flows into security intelligence. As data in corporation are very large, Big-data technologies (e.g. Map Reduce, HDFS, Spark, and etc.) give opportunities of storage and analytical processing of large structured and unstructured data sets. Data sources include website traffic logs, business processes logs and day-to-day transactional logs. Our approach is based on a rough idea, that the main data source of network supervising for the purpose of anomaly detection is transactional logs of whole corporation. Transaction algorithm loads each record from log files line by line and stores them in the same block if they were created in the same time division, within the same session, the IP addresses and ports are identical. Then it is decided whether this transaction fulfills the conditions to be included in the anomaly profile [6].

A rule always contains some basic attributes related to time and session parameters and also a device ID, from which particular log record originates. The anomaly finding algorithm sequentially processes the log files intended for analysis and creates transactions from these files. This transaction is used to identify an anomaly, and then it is stored for further observations.

Log dataset stream of events constantly fixed in transactional logs is unstructured big data, which require parallel storage and analytical processing. As corporation is multilevel hierarchical system, for network anomaly detection it would be better to consider hierarchical clustering of the overall network.

Phase I: Splitting

The input Log dataset stream can be split into several list of data sub-dataset.

Phase II: Mapping

In this phase data in each split is passed to a mapping function to produce output values. In our case, a job of mapping phase is to data clustering of each map from input splits. The K-means clustering algorithm consists of the following steps [3]:

Step 1. *K* initial cluster centers $z_1, z_2, \cdots, z_k$ are chosen randomly from the *n* points $x_1, x_2, \cdots, x_n$

Step 2. Assign point $x_i, \quad i = 1, 2, \cdots, n$ to cluster $c_j, j \in \{1, 2..., k\}$ iff (*if and only if*):

$$\left\| x_i - x_j \right\| < \left\| x_i - z_p \right\|, \quad p = 1, 2, \cdots, K \ \& \, j \neq p$$

Step 3. New cluster centers $z_1, z_2, \cdots, z_k$ are computed as follows:

$$z_j^* = \frac{1}{n_i} \sum_{x_i \in c_j} x_i, \quad i = 1, 2, \cdots, K$$

where n_i is the number of elements belonging to cluster c_j.

Step 4. If $z_i^* = z_i, \quad i = 1, 2, \cdots, K$ then terminate. Otherwise continue from Step 2.

After this phase we get an initial center for all predetermined clusters.

Phase III: Genetic algorithm

To obtain more accurate cluster, it is reasonable to combine the K-means algorithm and genetic algorithm into a hybrid algorithm [7]. It can search for the best cluster number k, then cluster after optimizing the k-centers. Number of clusters (k) is an important parameter to clustering quality. In order to obtain high-precision clustering results, a modified hybrid algorithm with genetic algorithm and clustering algorithm uses a special fitness function of genetic algorithm.

Algorithmically, the basic steps of GAs are outlined as below:

Step 1 [Population initialization]: random population of chromosomes is generated, that is, suitable solutions for the problem. Each individual represent a row-matrix 1× n where n is the number of observation, each gene contain integer [1, K] which represent the cluster which this observation belongs to.

Step 2 [Evaluation]: the fitness of each chromosome in the population is evaluated. Evaluate the desired objective function, where the task is to search for appropriate cluster classifications such that the fitness function is minimized. The clustering fitness function for the *K* clusters $c_1, c_2, \cdots, c_k$ is given by

$$f\left(c_1, c_2, \cdots, c_k\right) = \sum_{i=1}^{k} \sum_{x_j \in c_i} \left\| x_j - z_i \right\|.$$

Step 3 [New population]: a new population is created by repeating the following steps:

Selection: Select two parents (chromosomes) from a population according to their fitness value. The chance for each chromosome to be selected, as a parent, is determined according to its fitness. The aim of selection is to direct GA search towards promising regions in the search space. We employ roulette wheel selection in this study; where the individuals on each generation are selected for survival into the next generation according to a probability value. The probability of variable selection is proportional to its fitness value in the population, according to the formula below:

$$p(x) = \frac{f(x) - f_{\text{Min}}(\psi)}{\sum_{x \in \psi} \left\{ f(x) - f_{\text{Min}}(\psi) \right\}},$$

where *p* (*x*), selection probability of a string *x* in a population ψ and

$$f_{\text{Min}}(\psi) = \text{Min}\left\{ f(x) \mid x \in \psi \right\}.$$

2) Crossover: According to the crossover probability (Pc), new offspring (children) is generated from parents. In this step, we present a modified uniform crossover, which is done on each individual, in such a way that offspring is constructed by choosing the individual with a probability *Pc*.

3) Mutation: According to mutation probability (Pm), new offspring at each locus (position in chromosome) is mutated. For each individual, mutation operator is implemented as follows, first select two columns randomly from *ith* individual and then generate two new columns.

4) Accepting: new offspring is placed in the new population.

Step 4 [Replace]: Use new generated population for a further run of the algorithm.

Step 5 [Test]: If the end condition is satisfied, return the best solution in current population and stop. Step 6 [Loop]: Go to step 2.

Phase IV: Cluster evaluation

The determination of cluster quality is done by the entropy measures. The entropy is negative measure. The lower entropy means better clustering. The greater entropy means that the clustering is not good. The quantity of disorder is found by using entropy[8]. We consider entropy as an evaluation metric for our clustering. First, we need to compute the entropy of each cluster. To compute the entropy of a specific cluster, use:

$$H(i) = -\sum_{j \in K} p(i_j) log_2 p(i_j)$$

where p(ij) is the probability of a point in the cluster ii of being classified as class jj.

Similarly, we can compute the entropy of other clusters. So first, we need these probabilities of points for each cluster being classified as each class. Once you have the entropy of each cluster, the overall entropy is just the weighted sum of the entropies of each cluster. You can compute the overall entropy using the following formula:

$$H = \sum_{i \in C} H(i) \frac{N_i}{N}$$

where H is the entropy, Ni is the number of points in the cluster ii and N is the total number of points.

Phase V: Rough Clustering Analysis

The main resolution of clustering is to reduce the size and complexity of the dataset. In this paper, the clustering process is improved by cluster entropy estimation. Points that are not within a cluster become candidates to be considered anomalies. Outlier detection is an important task in clustering analysis. The outlier is identified by entropy, which indicates the degree of outlierness for every transaction in the Log dataset [8].

The meaning of cluster analysis is the elimination of clusters with lower entropy from further consideration. So, the remaining clusters go to the corresponding Reducers.

Phase VI: Reducing. The process continues at the Reducer phase. The first process is to collect all the data record from Reducers. Next, the system will merge the remain clusters in unified list, to be ranked by entropy descending.

Phase VII: Output the final results.

As final output we get a list identified of network anomalies for further decisions.

Conclusion

In this paper we presented an approach to network supervising for the purpose of anomaly detection in transactional logs of whole corporation. The anomaly finding algorithm sequentially processes the log files intended for analysis and creates transactions from these files. This transaction is then compared with the set of rules and if it is identified as an anomaly, it is stored for further observations.

Our approach is based on a multiphase algorithm, integrating the clustering based on K-means method with genetic algorithms and cluster entropy estimation, that has high performance quality of clustering the overall network. On the whole, the processing of the log file is implemented in Hadoop Mapreduce environment.

ინფორმატიკა

ქსელის ანომალიის გამოვლენა კორპორაციულ უსაფრთხოებაში

გ. ჯანელიძე*, ბ. მეფარიშვილი*, ლ. შონია*

**საქართველოს ტექნიკური უნივერსიტეტი, ინფორმატიკისა და მართვის სისტემების ფაკულტეტი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის გ. ჩოგოვაძის მიერ)

კორპორაციული უსაფრთხოების ხარისხი დამოკიდებულია ქსელის უსაფრთხოებაზე, რომელიც ერთ-ერთი მნიშვნელოვანი აქტივია კორპორაციულ უმრავლესობაში და კარგად უნდა იყოს დაცული ინფორმაციის უსაფრთხოების ეფექტური პრაქტიკის საშუალებით. ჩვენ ყურადღებას ვამახვილებთ კორპორაციული უსაფრთხოების გარკვეულ გამოწვევებზე, განსაკუთრებით იერარქიული კლასტერირების პრობლემაზე, როგორც ქსელის ანომალიის გამოვლენის პრობლემასა და შესაძლებლობებზე. K-means ალგორითმს, კლასტერისთვის გენეტიკურ ალგორითმთან ერთად აქვს კლასტერის მაღალი ხარისხი, პროცესის კონვერგენციის მინიმალური დროით. ჩვენ განვიხილავთ ენტროპიას, როგორც შეფასების მეტრულს ჩვენი კლასტერიზაციისათვის. იქიდან გამომდინარე, რომ მოწყობილობების, პროგრამებისა და მომხმარებლების მონაცემები ძალიან დიდია, Big-data ტექნოლოგიები იძლევა დიდი სტრუქტურირებული და არასტრუქტურირებული მონაცემების ნაკრებების შენახვისა და ანალიტიკური დამუშავების შესაძლებლობას. ჩვენი მიდგომა ემყარება იმ იდეას, რომ ქსელის ზედამხედველობის მონაცემთა ძირითადი წყარო ანომალიის გამოვლენის მიზნით, არის მთელი კორპორაციის აქტივობების ჩანაწერების ანალიზი. ტრანზაქციულ ჩანაწერებში მუდმივად დაფიქსირებული მოვლენების მონაცემთა ფაილი წარმოადგენს სტრუქტურირებულ დიდ მონაცემს, რომელიც მოითხოვს პარალელურ შენახვასა და ანალიზურ დამუშავებას Hadoop Mapreduce ფრეიმვორკის გამოყენებით. საბოლოო შედეგის სახით ვიღებთ ქსელის ანომალიების იდენტიფიცირებულ ჩამონათვალს შემდგომი გადაწყვეტილებების მისაღებად.

REFERENCES

1. Kemal A. Delic. (2018) Big data: corporate security is a big data problem. https://www.researchgate.net/publication/326645232
2. Ashish Bajpai et al. (2018) Big data analytics in Cyber Security, *International Journal of Computer Sciences and Engineering*. Open Access Review Paper, **6**(7), E-ISSN: 2347-2693.
3. Ahamed Al Malki, Mohamed M. Rizk, El-Shorbagy M.A., Mousa A.A. (2016) Hybrid Genetic Algorithm with K-Means for clustering problems, *Open Journal of Optimization,* 5:71-83, http://www.scirp.org/journal/ojop http://dx.doi.org/10.4236/ojop.2016.52009
4. Dhruba Kumar Bhattacharyya, Jugal Kumar Kalita (2014) Network anomaly detection: a machine learning perspective, Taylor & Francis Group, LLC CRC Press is an imprint of Taylor & Francis Group, an Informa business, ISBN-13: 978-1-4665-8209-5 (eBook).
5. Zeebaree D.Q., Haron H., Abdulazeez A.M. and Zeebaree S.R.M. (2017) Combination of K-Means clustering with Genetic Algorithm: a review, *International Journal of Applied Engineering Research,* **12**(24): 14238-14245. ISSN 0973-4562.
6. Kishan G. M., Chilukuri K. M., HuaMing H. (2017) Anomaly detection principles and algorithms, ISBN 978-3-319-67524-4 ISBN 978-3-319-67526-8 (eBook) https://doi.org/10.1007/978-3-319-67526-8
7. Xu W., Huang L., Fox A., Patterson D., Jordan M. (2008) Mining console logs for large-scale system problem detection, in workshop on Tackling Computer Problems with Machine Learning Techniques.
8. Chawla S. (2016) A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search, *Applied Soft Computing,* **46**:90-103. http://dx.doi.org/10.1016/j.asoc.2016.04.042

Received May, 2021