International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 02 || February 2017 || PP. 21-25

A survey of Big Data Analytics Techniques in Cyber Security

Ana-Maria Ghimeș [1], Victor Patriciu[1]
[1](Doctoral School, Military Technical Academy, Romania)

Abstract: In the time of huge information, there are a great deal of examination strategies and procedures for analyzing large data sets and acquiring applicable outcomes that are proposed to be used for specific purposes in various ranges of business.In the virtual environments, many attacks are launched for obtaining advantages through information leakages from their targets. The motivation behind investigation techniques in digital security is to end up distinctly more adaptable with changes in adversary behaviors. Visual examination and prediction algorithms seem to contribute considerable a lot in resolving cyber security issues. Exploring large data sets, achieving knowledge, forensic investigation, are representing the most known cases in cyber security big data solutions.To get significant information from analytics, the most important steps to take before analyzing data are to normalize, eliminate duplicates and put it in a format that can enhance the proficiency of an algorithm. Normalizing data is a pre-process that incorporate capacities and systems for sorting, mining, connection information and so forth.

Keywords: analytics, big data, cyber security, data mining

1. INTRODUCTION

Big Data is a term used to describe large data sets of structured and unstructured data which cannot be handled by the current tools and technologies. Big Data is characterized by seven features: volume, velocity, variety, veracity, value, variability, visualization.

The term of big data analytics is referring to the cycle of managing data: data gathering, data organization and data analytics to discover patterns, to understand states of the situations, to predict and to understand behaviors. There are a few stages to follow when you are working with big data:

- Data acquisitionrepresents the process of data gathering, filtering and normalizing data before it is stored in a data container or in a storing application.
- Data Analysis represents the process of raw data transformation in forming a decision.
- Data Maintenance represents the active management along a life-cycle to maintain the quality of analytic model and its utility. These stage includes actions like creation, selection, classification, transformation, validation and conservation of data.
- Data Storage should be made in scalable way maintaining all the requirements to get access to the data available.
- Data Utilization.

The main security issues in big data analytics are:

- Lack of control and transparency;
- Reuse of data;
- Data implications and re-identification;
- Data profiling and the decision making process.

Nowadays, big data analytics are representing the most effective defense against cyber threats. The analytics methods and techniques are used to detect intrusions, to predict behaviors and to make decision in case of security alert.

2. ANALYTICS METHODS

2.1 A/B Testing

The A/B Testing represents a technique used for comparing a control group with a variety of test groups for determining what changes/ treatments will improve an extension variable (e.g. the response rate in marketing). Additionally, this procedure is known as split testing or bucket testing. The A/B testing can be used to decide which copied text, image or colors could improve the transformation rates in a web based business webpage. Big Data permits executing and analyzing large number of group tests ensuring that these tests will be sufficiently large to detect the important differences between group control and comparisons groups. The A/B testing can be extended to A/B/N testing when multiple scope variables are needed to be used.

2.2 Association Rules

The Association Rules represents a set of techniques used for binding some relations between elements and variables from a large database (e.g. association rules). These techniques are applied in e-commerce environments, like analyzing shopping baskets for determining which products are purchased together. The data is then used for exposure and promoting. The set of techniques comprises from a variety of generated algorithms and from possible test rules. These are habitually used as a part of mining information. The most widely recognized uses of association rules are putting comparative items in a similar place to enhance sales, for extracting relevant information about customers from web servers, for analyzing biological results to discover new relations, for monitoring system logs, for discovering intrusion detection or malicious activity etc.

Dubey et al [4] is proposing in their paper a learning framework utilizing association rules and furthermore protecting privacy. The process of mining data (ARM) is composed from two parts. The first part comprises in generating a list of frequent items sets from all known items. To avoid generating a list with all the existing items, a threshold is set. This support value is filtering the item sets which can conduct to results that are not representing a certain scope. The second part comprises in generating the association rules from the list of frequent item sets. These rules are chosen about the support value (threshold). The two values, support and trust, are characterizing the amount you ought to trust the association rules produced through this process. In a multiparty model, the data sources can exist on multiple hard disks. There are two possibilities of processing data: parallel and distributed, and, usually, this implies coupled systems with shared memory (SMP), distributed memory machines or SMP's clusters.

Kaosar et al [10] are proposing an ARM technique for sure comparisons (correlations checks for a data set) and it is based on a homomorphic encryption scheme which implies reusing resources.

2.3 Classification trees

The statistic classification represents a computational model for identifying categories which are belonging to certain observations/ features. This method needs a training set of identified observations – historical data. The statistic classification is used in: automatically assignation of documents to certain categories, to classify organisms in groups, to develop student profiles which are following online courses.

2.4 Genetic algorithms

Genetic algorithms represent a computational model family which is based on the evolution and natural selection principles. These algorithms are transforming a problem from a certain domain in a model through chromosomal data structure usage and the chromosomes will evolve through selection, recombination and mutational operators. In the computer security, these algorithms are used for finding an optimal solution for a particular issue.

The genetic algorithms are inspired from how the evolution model works – through mechanisms like inheritance, mutation and natural selection. These mechanisms are used for observing how some available solutions of problems need to be optimized.

There are a lot of usages of genetic algorithms like scheduling doctors at emergency hospitals, returning some combinations of optimal materials or engineering work for designing efficient machines, etc.

The ways of a genetic algorithm usually starts with a chromosomal population randomly picked. These chromosomes are representations of the problem that need to be solved. In accordance with attributes of the problem, every positions of the chromosomes are encoded in bits, characters or numbers. The set of the chromosomes from an evolution stage is named population. An evaluation functions is used for computing the quality of every chromosome. During the evaluation processes, two operators, crossover and mutation, are used for simulating a natural reproduction and a mutation of species. The selection of chromosomes for survival and their combination is chosen based on the best chromosome.

The genetic algorithms can be used to develop simple rules for network traffic. Then, these rules are used for recognizing unapproved connections to the network. The unauthorized connections are referring at events with various probabilities of intrusion. A rule, usually, has this form:

If {condition} then { act}

Depending on the issues, the condition refers, usually, at a match between a connection to the current network and rules from an IDS, like source and destination IP addresses, port numbers, connection lifetime, protocol used etc., indicating the probability of an intrusion. The field act refers at the action defined by the security policies from an organization, like reporting system alerts to administrator, closing the connection, logging messages in an audit file, etc.

The purpose of using a genetic algorithm is to generate rules to match the abnormal connections. These rules are tested with the help of a connection history and are used to filter the new ones and to find malicious traffic in the network.

In different implementations, the network traffic used as input for the genetic algorithm is represented by pre-classified data to make distinction between authorized connections and the unauthorized ones. The data is gathered using network sniffers. The data collection is manually classified based on the expert knowledge. Running the genetic algorithm with a small data set of rules can generate a greater data set which will contain rules for IDS. These rules can be considered adequate for the algorithm and can be used for filtering the new network traffic.

A genetic algorithm is running with different parameters. Every parameter impact the adequacy of the algorithm. The evaluation function represents the most important parameter from the algorithm. The output of the function is computed based on the condition if exists a field that match the pre-classified data, and, then, it increases the yield with weight of that field. The order of the values is established according with various field from the connection (reported by sniffers). In this way, every gene which represents the IP destination address has a similar weight. The real values can be altered at the execution time. The main idea for ordering these values is the way that distinctive fields has diverse significance.

The absolute difference between the output of a chromosome and the level at a connection is considered suspicious is computed using the following formula:

$$\Delta = |\text{outcome-suspicious\_level}|$$

At the moment, a mismatch occurs, it is computed a penalty value. The ranking variable shows the level at which an intrusion is easily or not identified:

$$\text{penalty}=((\Delta*\text{ranking})/100)$$

The matching of a chromosome is computed taking into consideration the penalty mentioned above:

$$\text{fitness}=1\text{-penalty}$$

When a genetic algorithm is used, it is necessary to set a local maximum (acceptable solution) different from global maximum (the best solution). The niching techniques can be used to find some local maximums. These techniques are based on the analogy with the nature where every environment are composed by many subspaces which support different types of life. The genetic algorithm maintains diversity for every population in a multimodal domain. There are two types of methods to keep diversity in a population: crowding and sharing. The crowding method is using similar members for replacing and slowing down the population to converge to a single point in the next generations. The sharing method reduces the conformity of some individuals which have similar members and are forcing these members to progress through different local maximums which are less populated. The use metrics for similarity in these techniques could be phenotypes for genotyp similarity like Hamming distance between representations of bits, or similarity phenotype like relation between two connections.

Folino et al [6] used the genetic programming to build an engine named CelluarGenetic programming (CAgE), which runs on shared memory computers and in distributed environments using fine-grained model. The generation of a population and different operators are defined like in the classic genetic programming. Comparing with the classic models, in their approach, the classifiers represent leaves of the tree, and the combining functions represent the nodes. The genetic programming is used for developing a combining function, which will then be used by assemblies for adopting new sets of classifiers. The function selects the most eligible models/ classifiers for a specific data set. Particularly, the chosens functions for combining classifiers are functions that cannot be trained, like average, weight average, multiplication, maximum and median. These can be applied on a various number of classifiers. Every function is multiplied with different parameters (arity). The only function that is not included in this set is the product.

2.5 Machine learning

This method includes software that can learn from data. It offers computers the ability to learn without being explicitly programmed, and it concentrates on making predictions based on features learned from training sets. The machine learning techniques are used to differentiate between spammers and non-spammers emails, to learn preferences of users and to make recommendations based on these information, to determine the best content for the clients, to compute probabilities of winning a case and to set different legal rates.

In their paper, Mehmood et al al. [12] summarize all the anomaly detection systems that can be implemented using machine learning techniques:

- Fuzzy Logic, which uses true or false values for detecting anomalies;
- Neural Networks, which compares different inputs and it transforms them until the needed result is obtained;
- Support Vector Machines, which classifies the normalized data using different kernels for

splitting data in two categories resulting in the anticipation of the new inputs;

- K-Means, which classifies data in clusters and then computes a media of these using different techniques.

Anna L.Buczak et al [1] used machine learning for generating a data set for training a test collection data from penetration tests. The features for detection models of tunneling were determined in an iterative mode. The initial iterations were based on literature and analytics specifications and the last ones were based on source data analyzes. In their approach, they used random forest classifiers, a machine learning technique which combines decision trees with bagging trees. The forest is composed from different decision trees and the final prediction is determined by the majority of electors. Every decision tree is formed by random features. The combination between decision tree and bagging tree is used for reducing the variance of a static machine learning method. This process starts generating from a training set with dimensions n and m and continues with training sets of dimension n through levelling and replacement. In this approach, there are used only two tierce from all features. The remaining features are used for computing the error rate. In this paper, the method proposed are describing a forest trained to distinguish between normal activities and tunneling.

2.6 Regression analysis

The regression analysis implies the manipulation of various independent variables for observing how they are influencing a dependent variable (e.g. the time spent in a shop). This analysis describes how the value of a dependent variable is changing depending on the variance of an independent analysis. It is used to find how the satisfaction level affects the loyalty of the clients, how the support calls may influence the prediction of the weather from the former day etc.

The linear regression represents a tool for statistical analysis in the data mining process, because it permits to control the dependency between a dependent variable and an independent one. The fully homomorphic encryption offers a solution for computing statistical analysis over the encrypted data and, in the same time, is preserving and protecting privacy.

Fang et al [5] are addressing the compromise between confidentiality and statically analysis using linear regression. To keep the compromise at a low level, they are using PPRCP (Privacy Preserving Regression Coefficient Protocol). This protocol is computing the regression coefficient without revealing sensitive data. The data owners are exposing only the regression parameter and the error detection model.

2.7 Sentiment analysis

The Sentiment analysis helps researchers to determine the speaker/ writer feelings about a certain subject. This technique is used usually to enhance services in a hotel through analysis of clients’ comments, to customize services to address the client issues, to determine the real opinions of the client in social media.

2.8 Social network analysis

Social network analysis is a techniques used from the beginning in the telecommunication industry, and then it was quickly adopted by sociologists to study interpersonal relationships. It is used to study relations between individuals from different domains and commercial activities. Most of time, it is represented through a graph, where the nodes are the individuals, and the edges are representing the relations between them.

Social network analysis is used usually to study how people from different areas are connecting with each others, to find the importance or the influence of an individual in a group, to find the minimum relations that are connection two individuals, to understand the social structure based on a client etc.

The analysis of Twitter messages as security alerts sources represents a very good example of social network analysis usage. These messages were analyzed to verify it they indicate potential security issues in virtual environments. It was formulated two hypothesis [2]: there is information about computer’s security in tweets and some of them are indicating potential threats and Twitter are reporting security issues before specialized sites. The data is gathered through a specialized software as the proposed approach by Campiolo et. al [2], then the data is filtered keeping only the messages that contain security data, the next step is to index and organize it in groups. The last stage is to make correlations between tweets and news. In their paper, Campiolo et al [2] concluded that 60% messages are representing security alerts, 31% are security messages and 9% are spams or false-positive messages.

3. CONCLUSIONS

After reviewing different analytic methods and different implementation of them for resolving cyber security issues we can state that there is not an optimal and unique solution for data mining and big data in security information. Big Data is still a new technology which tends to be the base for a great variety of applications. The applications which use Big Data technologies will improve fast risk management systems through different approaches of analytic techniques. Data analysis and data mining techniques aid on creating user models, based on collected data. User profiling, risk management systems, intrusion detection systems are all applications that can use analytical methods behind.

REFERENCES

[1]. Anna L. Buczak, P. A. (2016). Detection of Tunnels in PCAP Data by Random Forests. *CISRC '16: Proceedings of the 11th Annual Cyber and Information Security Research Conference.*

[2]. Campiolo, R., Santos, L. A., Batista, D. M., & Gerosa, M. A. (n.d.). Evaluating the Utilization of Twitter Messages., (p. 2).

[3]. CHOR, B., KUSHILEVITZ, E., GOLDREICH, O., & SUDAN, M. (1988). Private information retrieval. *Journal of the ACM*, 965-981.

[4]. Dubey, P., & Dubey, R. (2014). FULLY HOMOMORPHIC ENCRYPTION BASED MULTIPARTY ASSOCIATION RULE MINING. *International Journal of Computer Engineering & Science*, 14-17.

[5]. FANG, W. Z. (2012). Privacy Preserving linear regression modeling of distributed databases. *Optimization Letters*, 807-818.

[6]. Folino, G., Pisani, F. S., & Sabatino, P. (2016). An Incremental Ensemble Evolved by using Genetic Programming to Efficiently Detect Drifts in Cyber Security Datasets. *GECCO '16 Companion Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, (pp. 1103-1110). Denver, Colorado.

[7]. Goldwasse, R. B. (2014). *https://eprint.iacr.org/2014/331.pdf.* Retrieved from https://eprint.iacr.org: https://eprint.iacr.org

[8]. Joppe W. Bos, K. L. (2013). Private predictive analysis on encrypted medical data. *Microsoft Tech Report 200652*.

[9]. Kantarcioglu, M., & Xi, B. (2016). *CCS '16 Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1866-1867). Vienna: ACM.

[10]. KAOSAR, M. G., PAULET, R., & YI, X. (2012). Fully homomorphic encryption based two-party association rule mining. *Data & Knowledge Engineering*, 1-15.

[11]. N. N. Dalvi, P. M. (2004, August 22-25). Adversarial classification. *10th ACM SIGKDD*, pp. 99-108.

[12]. Y. Mehmood, M. A. (2013). Intrusion Detection System in Cloud Computing:. *2nd National Conference*, (pp. 59-66). Rawalpindi, Pakistan.