Big Data Feature Selection Model for Intrusion Detection using Data Analytics

Dr. Sai Prasad Padavala[1]
Assoc. Prof., Dept. of Computer Science and Engineering,
Sri Venkateswara College of Engineering & Technology (Autonomous),
Ranked 176 in INDIA By NIRF 2019-MHRD, Chittoor, A.P, INDIA.

Sangeetha Tamilarasu[2]
Research Scholar, School of Computer Science and Engineering,
Vellore Institute of Technology, Vellore, TAMIL NADU,INDIA.

Samatha Gurava [3]
Asst. Prof., Dept. of Computer Science and Engineering,
Sri Venkateswara College of Engineering & Technology (Autonomous),
Ranked 176 in INDIA By NIRF 2019-MHRD, Chittoor, A.P, INDIA.

***ABSTRACT*: Today a huge amount of data is present in the internet as the network technology and the information technology has been developing rapidly. With the lack of knowledge on threats and security providing sources for the big data, it has been a big challenge for the Intrusion detection. It is a time taken process to search for the related features in the intrusion detection system which is a big challenge in developing big data in it. In this paper, big data feature selection model for intrusion detection using data analytics is proposed. First, features search space is converted into binary vector from a continuous vector space by using a sigmoid function to be suitable for the feature selection problem. The random initialization of the parameters for the algorithms can be achieved by implementing the tent chaotic maps. Then the behavioral and content features are generated for analyze the characteristics of network traffic and information present on payload. The features selection can be done in parallel using the k-means clustering and big data methods with deployment of multiple machines in the network. Then deep learning based FCN, CNN, and RNN classifiers are used to train the model in parallel. This effectively reduces the time taken to build the proposed model. The proposed model achieves better accuracy of classification and intrusive attack detection rates can be increased by using this model in comparison with the other earlier model approaches.**

KEYWORDS: Big data, Intrusion Detection System (IDS), feature selection, deep learning.

I. INTRODUCTION

With the greater improvement in the utilization of internet networks along with the fast increase in the amount of data originating from various sources, the most significant challenge to provide protection and privacy for big data has been faced by the many of security management system's developers. One of the most significant and commonly used systems in the cyber security systems are the Intrusion detection systems [1]. The method of compromising the integrity, confidentiality and availability of resources in the security mechanisms among the networks or computers is referred as intrusion detection systems. They can be either hardware or software which monitor as well as analyze the flow of data through the networks or computers for detecting the security violations that intimidate integrity, confidentiality or availability of resources in system [2]. However, a difficult challenge faced by these intrusion detection systems is to monitor and analyze the heterogeneous data. From the past few decades, several big data managing and computing technologies have been created including NoSQL, Hive, Hadoop and Apache Spark. Various advantages have been offered by the big data techniques that include processing storing and receiving speed for various data types.

Thus to address the challenges of intrusion detection systems, big data integrated with deep learning techniques are used for compromising the security management and computation with the help of Apache Spark technology and the Keras deep learning library. The degree of data homogeneity can be calculated for the feature selection using the k-means clustering approach. In prior such features are classified using the deep learning based classifiers for each cluster sample that include Fully Connected Networks (FCN), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). These models are applied to the data set of UNSW-NB15 which consists of a normal and malware network attack samples [3] [4].

II. FEATURE SELECTION IN IDS

The most considerable and useful system that has been protected the computers or networks from the unknown attacks is the intrusion detection system [5]. The signature based and anomaly based detections are the two ways of intrusion detection system used to detect the attacks [6]. The machine attacks are determined using the signature-based analysis by detecting the attack signatures against the recognized database attack signatures while the attacks in the machine based way on the behavior are detected using the anomaly based detection by monitoring and alerting the issues against a normal baseline. Generally, a huge amount of data which includes various types of traffic patterns has been processed by IDS. The set of attributes or features characterize the every pattern of dataset then a point is represented in a multi dimensional feature space. Features that are redundant and irrelevant might be contained in a pattern with which the process of training and testing can slowdown or even the performance of classification can affected by the grater mathematical complexity. On the other hand, in fact, it is worth keeping smallest possible number of features to reduce the structure complexity and computational costs of a classifier. Elimination of insignificant features also provides the data visualization capabilities, improved modeling methods, performance of prediction, and fast classification process. Therefore to address this problem, the feature extraction and feature selection of dimensionality reduction methods are applied successfully to data mining and machine learning models.

The input applied features set is transferred into a new set of features using the Feature extraction methods. Then the features that are more informative are selected by the searching from the original input data using the Feature Selection (FS) algorithms [7].

Filter, wrapper and hybrid are the three categories of feature selection methods in which they are generally classified. On the one hand, relationship between the group of features is estimated by a criterion of using an independent measures such as distance, information, or consistency in the filter algorithm by searching the features starting from a vacant subset. This process of searching carried out till reaching of a preferred number or until producing a no better feature subset by adding or deleting of any feature. This filter algorithm output is a best optimum feature subset. Computationally this method has treated as less expensive also it has been argued that this method can applicable easily to the high dimensional datasets and is more used in general. But, since the filter methods failed in choosing best optimum feature subset and sometimes may also select the redundant features because of the no interaction present between the classifiers and feature's correlation, it results may not good enough always [8]. Therefore, performance of deep learning based classification models is varied depending on such selected features also it greatly depends on the quality of criterion used for selection.

Then a specific learning algorithm such as support Vector Machine (SVM) or Decision can be used in the wrapper algorithm method as a fitness function to estimate the features integrity. This method has been treated as a more accurate one than the filter method but has generally more computational complexity compared to filter method when large feature subsets are used [9]. After that, the two concepts and their advantages of filter method as well as wrapper method are concentrated by the many of researches in order to address the abovementioned limitations and introduced a hybrid algorithm. It utilizes both the independent measures and fitness evaluation function of feature subset. The final best optimum feature subset is selected with this hybrid algorithm by using the concept conveyed in filter algorithm along with the specific machine learning [10]. The better performance can be achieved by this method with more effectiveness however it may not as fast as that of the filter method.

III. BIG DATA FEATURE SELECTION MODEL FOR INTRUSION DETECTION

Figure (1) shows the flow diagram of the proposed big data feature selection model for the intrusion detection using data analytics. This model primarily consists of initialization, feature generation, deep learning based classification, evaluation, transformation and iterations.

3.1 Parameter Initialization

Initialization of n number of search engines or wolves population that are generated randomly is the first step. A desirable solution is represented by the each search agent related to the dataset and 'd' is the length corresponds to the number of features in that dataset. Some features which improve the classification accuracy required to be selected in order to summarize the feature selection problem of the purity classification. Thus, it is necessary to consider the related features defined by one value while ignoring the other features defined by zero. Initially, each solution was configured with the 1's and 0's (i.e. with binary values).

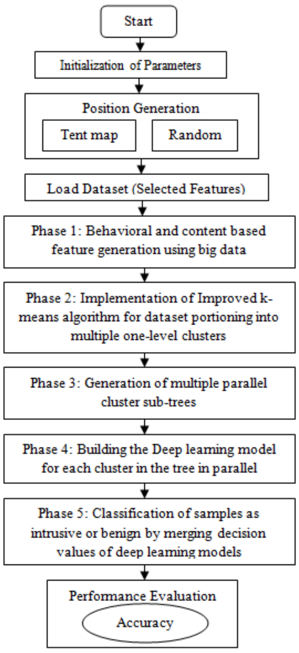
Start
Initialization of Parameters
Position Generation
Tent map
Random
Load Dataset (Selected Features)
Phase 1: Behavioral and content based feature generation using big data
Phase 2: Implementation of Improved k-means algorithm for dataset portioning into multiple one-level clusters
Phase 3: Generation of multiple parallel cluster sub-trees
Phase 4: Building the Deep learning model for each cluster in the tree in parallel
Phase 5: Classification of samples as intrusive or benign by merging decision values of deep learning models
Performance Evaluation
Accuracy

Fig. 1: FLOW DIAGRAM OF PROPOSED BIG DATA FEATURE SELECTION MODEL

Various random initialization methods such as Distributed Sampling (DS) based, chaotic map based, etc., are there for initialization of parameters. Instead of using logistic map method, tent map can enhance the homogeneity of transverse and algorithm optimization can also speed up in between a better value of [0,1] . The proposed algorithm's convergence property depends on the random sequence of numbers that are utilized to execute the algorithm with different parameters. The mathematical model of tent map is defined in the following eq. (1).

$$x_{k+1} = \begin{cases} \mu x_k & for\ 0 \leq x_k \leq \frac{1}{2} \\ \mu(1 - x_k) & for\ \frac{1}{2} \leq x_k \leq 1 \end{cases} \text{ --- (1)}$$

R is the random value in between [0, 1] and 1 + R is set to the µ of the above eq. (1). From the above discussion, every point of x_0 is generated after modification of procedure in the interval in [0,1] using x_k.

3.2 Updating the Position

In addition with the position information of third best wolf packs solution, the position of some best and second best wolves are taken into account by the Gray Wolf Optimization (GWO) algorithm in this position updating step which allows the sharing of learning content between a certain and pack of wolves. However the transfer of information between a person and their own knowledge might ignore. Therefore, it implemented the (Particle Swarm Optimization) PSO algorithm concept in order to improve the process of position updating. This PSO algorithm changes the present position of the particle using best position data of their own particle and group of particles. In conjunction with the GWO equation defined by position updating, the PSO algorithm thought to use the experience of individuals in this paper. Therefore, following gives the new position update equation.

$$x_{i(t+1)} = c_1 r_1 (w_1 X_1(\mathrm{t}) + w_2 X_2(\mathrm{t}) + w_3 X_3(\mathrm{t})) + c_2 r_2 (X_{\mathrm{ibest}} - X_{\mathrm{i}}(\mathrm{t})) \text{ ------ } (2)$$

Here, r_1 and r_2 represent the random variable set between the [0, 1] range and c_1 and c_2 represent the social and cognitive learning factors respectively. Best position of the gray wolf is denoted by the X_{ibest} set. The set of inertia weight coefficients are w_1, w_2, w_3 that can be measured by the following equations.

$$w_1 = \frac{|x_1|}{|x_1+x_2+x_3|} \text{ ------- } (3)$$

$$w_2 = \frac{|x_2|}{|x_1+x_2+x_3|} \text{ ------- } (4)$$

$$w_3 = \frac{|x_3|}{|x_1+x_2+x_3|} \text{ ------- } (5)$$

Since the nonlinear control parameters are proved as better one than linear optimization approach, it is used in this and expressions is given by,

$$a_1(t) = a_s - (a_s - a_f) \times \left(\frac{t}{tmax}\right)^2 \text{ --- } (6)$$

Here, a_s signifies the starting parameter of control and a_f signifies the end parameter value of control. Then the t and $tmax$ groups indicates the latest and the overall iterations number of algorithm.

Algorithm generates continuous position values for search agents. This can't be directly applied to the proposed model since it faces the binary format standard feature selection challenge. The given classification algorithm's accuracy and performance is improved by selecting the more suitable features based on feature selection problem. The transform function is used to transform the calculated or resultant search space from continuous to a binary format. The S-shaped function is determined by using the sigmoidal function. Following sigmoid functions can be used to translate every continuous value into binary value.

$$x_i = \frac{x_i - Min}{Max - -Min} \text{ ------- } (7)$$

$$x_{si} = \frac{1}{1+e^{-x_i}} \text{ ------ } (8)$$

$$x_{binary} = \begin{cases} 0 & if\ \ R < x_{si} \\ 1 & if\ \ R \geq x_{si} \end{cases} \text{ ------ } (9)$$

Here, Max represents the maximum value of continuous feature vector and Min represents minimum value of continuous feature vector. x_{si} is the continuous feature value of search agent in the S-shaped search space. $i = 1, \cdots, d$ and the value of x_{binary} determined by random number $R \in [0,1]$ can be either 0 or 1 in comparison to the x_{si}.

3.3 Generation of behavioral and content features

In any machine learning based models, the foremost significant operations are the generating and extracting of features due to the significant impact of several feature selection types on the performance of machine learning models. This section describes the generating procedure of behavioral and content features. First the generation of behavioral features is carried out when the network traffic analyzation was completed. In this model, Behavioral features consist of used protocol type, several statistics of steam level or packet level switching, collecting duration and the number of bytes of data transferred from source to destination. Then generation of content features is carried out after completion of behavioral features generation. The main location for initiating any type of malicious attacks, exploiting of illegal methods or commands is the payload in which most important information is embedded. This most important information is exposed by the extraction of content features from such payloads. Then one of the most efficient methods to select the content features is to use the measure of information gain as its selection criterion. The samples can be most effectively classified by the high information gained feature. The Spark framework based memory Map Reduce method is used in cluster of machine for performing the parallel extraction and selection tasks of byte n-gram to efficiently and effectively provide solution for feature selection.

3.4 Generation of one-level clusters

The one-level clusters are generated using the K-means algorithm with the generated behavioral features and content features. The development of Apache Spark along with the parallel improved K-means clustering is used in this to address the problems of scalability and random selection. The random initialization limitation can be overcome with this improved K-means algorithm by adopting the greedy initialization techniques. The final cluster solution that represents the more accurate and consistent data distribution can result by choosing the appropriate staring points which can be done by the greed initialization technique

The mapping and reducing are the two phases include in the parallel k-means clustering approach for the all iterations. The nearest centroid is calculated in a parallel way for each sample placed in different data partitions in the mapping phase. Different machines are used to place the every data partition. Then after obtaining of partial sum from the every cluster regarding to each dataset partition the centroid is re-calculated for each cluster in the reducing phase. Handling of millions of samples can be enabled by the clustering algorithm efficiently with such strategy of divide and conquer. Following eq. (10), defines the distance score between the sample and the cluster center.

$$dist(x, c) = \sum_{i=1}^{N} |F_x(i) - F_c(i)| \text{ ---- } (10)$$

Here, N signifies the number of features for each sample. $F_x(i)$ and $F_c(i)$ are the feature values of sample x and given cluster's centroid c respectively at an index i.

3.5 Building of Deep learning model for each cluster

In the multilevel hierarchical tree structure, every cluster still includes the irrelevant and noisy information as the accurate defining of distance function is difficult for the calculation of similarities among the samples. The intrusion detection system effectiveness can be greatly reduced with a considerable amount of such irrelevant and noisy information. According to the combined behavioral and content features, every cluster of the multilevel cluster tree is trained with deep learning model to enhance the performance of intrusion detection system in the hierarchical tree structure. In this multilevel cluster tree structure, best deep learning model is selected for each cluster by the evaluation Fully Connected Network (FCN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) of three different models. The distinctive information of each cluster in a traffic sample payload and the network traffic property patterns are analyzed by the deep learning model which designed through the combination of behavioral and content features.

Very similar samples of every cluster are concentrated with no distraction from the other clusters unrelated information by the each deep learning model. Therefore each deep learning model can achieve improved efficiency. Mapping the function for the more number of layers to the output vector from the input vector is the main objective of deep learning based FCN, CNN and RNN. Mapping function is defined in the following Eq. (11) defines for the each layer of every deep learning model.

$$a_{i+1} = f_i(W_i \times a_i + b_i) \text{------ (11)}$$

Here, f_i signifies the activation function while W_i, a_i, and b_i represents the weight matrix, activation factor and bias of the ith layer. The difference between predicted output and actual output is calculated using the loss function. Thus for such three types of proposed neural networks this loss function can be defined as,

$$Loss(y, \hat{y}) = \sum_{i=1}^{M}[y_i \log \hat{y} + (1 - y_i) log \hat{y}] \text{ ------ (12)}$$

Here, y_i represents the true label for sample i and $\hat{y}$ represents the deep learning model output for sample i. M signifies the each cluster size. y_i can be represent in the range between [0,1], from which 1 indicates a malware and 0 indicates benign sample.

3.6 Decision Fusion Algorithm

The following Eq. (13) expressed the decision function of deep learning model for k cluster to classify x sample as intrusive or benign.

$$f_k(x) = Sigmoid(W_k \times y_k + b_k) \text{ --- (13)}$$

Here, y_k and W_k represents the last layer activation and weight matrix in between end and output layer of k^{th} cluster in the neural network respectively.

The performance of deep learning models in predicting the samples as intrusive or benign can be evaluated using the decision values. The deep learning based models with very different data distributions can be build in different clusters. The comparison between values of deep learning model classification and various models can be carried out objectively using the z-score by normalizing the deep learning model's classification value $f_k(x)$. Since the calculation of deep learning models decision function for various clusters was done by the samples related to the same cluster which forms different high dimensional feature spaces, this process of normalization is required. The following Eq. (11) express the normalized decision value of sample x for the cluster k

$$decision\ value_k(x) = \frac{f_k(x) - mean_k}{\delta_k} \text{ --- (11)}$$

where $mean_k$ is the mean classification values of the deep learning model in cluster k and δ_k is the standard deviation of classification values of the deep learning model for the cluster k. Higher decision value shows that the deep learning model is more confident to classify a given sample.

3.7 Evaluation

If more than one aim is there for the problem to solve and get best solution then the feature selection problem is called as multi objective problem. According to this, problem of best features selection is considered as the "multi objective problem" although it needs to be done the following:

1. Have to select the reduced or minimum features of the dataset.
2. Have to select the classifier with improved or maximized performance and accuracy.

IV. RESULTS

In this section Apache Spark is used to test the proposed deep learning models of FCN, CNN and RNN classifiers. The process of classification carried out using more than one classifier is known as multiclass classification. Thus a dataset called UNSW-NB15 is used for the evaluation in which data packets representing the normal network traffic are not considered and the data packets representing the nine different types of attacks are considered. Then the FCN, CNN and DNN are tested using this dataset and shows the results of classification in Table (1).

Table 1: RESULTS OF CLASSIFICATION WITH UNSWNB15 DATASET

Classifier	Accuracy	Prediction
FCN	97.02%	5.05%
CNN	94.63%	15.67%
RNN	91.75%	23.96%

In comparison with the conventional techniques, the time of training and prediction for the proposed three classification models is enhanced significantly with the Apache Spark technique. This enhancement provides ability for the intrusion detection systems in making most effective and efficient decisions as related to the data allowing or blocking in the network. The deep learning models ability can be increased most effectively and very quickly with the Keras Deep Learning Library in addition to the integration of Apache Spark. In order to address the most important problem of intrusion detection system, the feature selection model proposed in this paper offers some improved performance for the classifiers. A better accuracy rate of 97% is obtained on dataset of UNSW-NB15 with the proposed method.

Table 2: COMPATIVE ANALYSIS ON ACCURACY FOR DIFFERENT METHODS

Methods	Classifier	Accuracy
[7]	RapTree	79.20%
	Random Tree	76.21%
	Naïve Bayes	73.86%
	Artificial Neural Networks	78.14%
[8]	Genetic + SVM	86.96%
Proposed method	FCN	97.02%
	CNN	91.75%
	RNN	92.34%

The comparative analysis on accuracy for different methods along with the proposed model is depicted in Table (2). These perform the multiclass classification on UNSW NB15 dataset.

V. CONCLUSION

In this paper, deep learning models integrated with the big data methods are developed in the intrusion detection systems in order to enhance their performance. The behavioral features and content features are generated using the big data methods. Then the best optimum feature subsets are classified and estimated by using the three deep leaning models of FCN, CNN and RNN classifiers. Based on these selected features, attacks are detected by using k-means clustering method in multiclass model. The proposed learning classifiers are tested on the dataset of UNSW-NB15 using an Apache Spark Technique with Keras Deep Learning Library. High accuracy levels are obtained in the results with our proposed models for detecting the attacks in multiclass classification with low prediction times compared to the other earlier studies.

VI. REFERENCES

[1] Junho Hong and Chen-Ching Liu, "Intelligent Electronic Devices With Collaborative Intrusion Detection Systems", IEEE Transactions on Smart Grid, Volume: 10, Issue: 1, 2019.

[2] Hamidreza Sadreazami, Arash Mohammadi, Amir Asif and Konstantinos N. Plataniotis," Distributed-Graph-Based Statistical Approach for Intrusion Detection in Cyber-Physical Systems", IEEE Transactions on Signal and Information Processing over Networks, 2018

[3] J. Slay and N. Moustafa, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set", Information Security Journal: A Global Perspective, 25(1-3), pp. 18-31, 2018.

[4] C. Yin, Y. Zhu, J. Fei and X. He, "A deep learning approach for intrusion detection using recurrent neural networks", *IEEE Access*, vol. 5, pp. 21954-21961, 2017.

[5] R. Wald, T. M. Khoshgoftaar and R. Zuech, "Intrusion Detection and Big Heterogeneous Data: A Survey", Journal of Big Data, 2(3), pp. 1-41, 2015.

[6] C. W. Ten, J. Hong and C. C. Liu, "Anomaly detection for cybersecurity of the substations", *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 865-873, Dec. 2011.

[7] Khalil El-Khatib, "Impact of Feature Reduction on the Efficiency of Wireless Intrusion Detection Systems", IEEE Transactions on Parallel and Distributed Systems, Volume: 21, Issue: 8, 2010

[8] J. Jiang, Z. Wu and Y. Peng, "A novel feature selection approach for biomedical data classification," Journal of Biomedical Informatics, volume: 43, number: 1, pp: 15–23, 2010.

[9] X. Xu, J. Huang and Y. Cai, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," Pattern Recognition Letters, volume: 28, number: 13, pp. 1825–1844, 2007.

[10] L. Yu and H. Liu, "Toward integrating feature selection algorithms for classification and clustering," Knowledge and Data Engineering, IEEE Trans, volume: 17, number: 4, pp. 491–502, 2005.

[11] Dong Song, M.I. Heywood, A.N. Zincir-Heywood," Training genetic programming on half a million patterns: an example from anomaly detection", IEEE Transactions on Evolutionary Computation, Volume: 9, Issue: 3, 2005

[12] Bo Sun, Kui Wu and U.W. Pooch, "Towards adaptive intrusion detection in mobile ad hoc networks", IEEE Global Telecommunications Conference, 2004. GLOBECOM '04., Volume: 6, 2004

[13] M. I. Heywood and A. N. Zincir-Heywood, "Dynamic page-based linear genetic programming", *IEEE Trans. Syst. Man Cybern. B. Cybern.*, vol. 32, no. 3, pp. 380-388, 2002.

[14] L. Wenke, S. J. Stolfo and K. W. Mok, "A data mining framework for building intrusion detection models", *Proc. IEEE Symp. Security Privacy*, pp. 120-132, 1999.

[15] R. Battiti, "Using mutual information for selecting features in supervised neural net learning", *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537-550, Jul. 1994.

Dr. Sai Prasad Padavala [1]

[1]. Assoc. Prof., Dept. of Computer Science and Engineering, Sri Venkateswara College of Engineering & Technology (Autonomous), Ranked 176 in INDIA By NIRF 2019-MHRD, Chittoor, A.P, INDIA.

Sangeetha Tamilarasu [2]

[2]. Research Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TAMIL NADU,INDIA.

Samatha Gurava [3]

[3]. Asst. Prof., Dept. of Computer Science and Engineering, Sri Venkateswara College of Engineering & Technology (Autonomous), Ranked 176 in INDIA By NIRF 2019-MHRD, Chittoor, A.P, INDIA.