IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 08-12
www.iosrjournals.org

The Use of K-NN and Bees Algorithm for Big Data Intrusion Detection System

Saqr Mohammed Almansob[1], Aqueel Ahmed Jalil[2],
Dr. Santosh Shivajirao Lomte[3]
[1](Department, of Computer Science, BAMU, Aurangabad, M.S India)
[2](Department of Computer Science, BAMU, Aurangabad, M.S India)
[3](Department of Computer Engineering, BAMU, Aurangabad, M.S India)

Abstract: *Big data problem in intrusion detection system is mainly due to the large volume of the data. The dimension of the original data is 41. Some of the feature of original data are unnecessary. In this process, the volume of data has expanded into hundreds and thousands of gigabytes(GB) of information. The dimension span of data and volume can be reduced and the system is enhanced by using K-NN and BA. The reduction ratio of K-DD datasets and processing speed is very slow so the data has been reduced for extracting features by Bees Algorithm (AB) and use K-nearest neighbors as classification (KNN). So, the KDD99 datasets applied in the experiments with significant features. The results have gave higher detection and accuracy rate as well as reduced false positive rate.*

Keywords: *Big Data; Intrusion detection system; K-NN method; Bees algorithm; feature selection*

I. Introduction

Nowadays, security has become very important for all computer networks in which big data are stored. For this reason, intrusion detection system monitoring and analysis all the data to detect intrusion detection over a network. The traffic that goes through it and significant component in network security. Now the use of information has expanded tremendously. There are hundreds and thousands and even millions of bytes of which data is being stored. There are gigabytes which are stored in a large quantity. The storage of data is made in the data storing devices, what is more, important than the storage of data is the retrieval of data, the classification of data, feature selection of data is necessary. Along with these the detection of data is an important task which needs to be carried and in this the other must important factor is that of securing. In all computer nodes, the mechanism related to security has become quite essential.

The next section will explore the related work. The section three discribes the proposed work. The section fourth show the experment and results. The final section It includes the conclusion and reference.

II. Related work

A few years back, there are various methods applied to feature selection. Rung-Ching Chen et al. [1] have proposed Rough Set Theory (RST) and Support Vector Machine (SVM) for intrusion detection. The authors used RST to preprocess and reduce high dimensionality of data. Furthermore, used SVM as a classification method. They used KDD data for experimented.. Anirut Suebsing et al. [2] Applied Euclidean distance-based feature selection and Cs.0 Algorithm for detect intruders over a network. The authors utilized feature selection technique for choose features which have high information to build a model for improving performance production. furthermore, used Cs.0 Algorithm as a classifier. The obtained results give improve a performance of detection rate based Euclidean distance. Kazeem et al. [3] have introduced Membrane Computing paradigm tool and feature selection method for improving Bees Algorithm(BA). They applied KDD-cup dataset in the experiments. so, the authors noticed that Membrane Computing very useful tool for rising classification accuracy rate and reducing the false alarm rate. Adel Sabry et al. [4] have proposed ID3 and Bees Algorithm(BA) for intrusion detection. They applied Bees algorithm to choose and generate best features for intrusion detection system(IDS). Furthermore, used ID3 as a classifier. So, they utilized KDD cup99 in experiments. The authors observant the proposed model ID3-BA give increased accuracy rate and decreased false positive rate. Abdullah et al [5] Applied Support Vector Machine (SVM) and K-Nearest Neighbor for intrusion detection problem. The authors applied PSO, Meta-optimized PSO and WMA techniques for enhancing accuracy rate and reduce false positive rate. so, the results obtained show the PSO approach give best accuracy rate Compared with the Meta-optimized PSO and WMA approaches. Zhengyu Deng [6] have proposed K-Nearest Neighbor algorithm (K-NN) for addressing and classification big data problem. They applied K-means clustering to distribute the big data into several parts. Furthermore, work on training and testing phases. The result obtained give higher accuracy and efficiency when using the K-NN algorithm for big data classification. Tingwel et al [7] have proposed Latent

Digichat Allocation (LDA) for discover patterns from big data in intrusion detection. The authers solved the problem of identifies the hidden pattern from big data by using LDA approach. So, intrusion detection problem mapped into topic modeling problem.

III. The proposed work

3.1. K-NN algorithm

The K-nearest neighbor (k-NN) is supervisor machine learning technique for object classification. K-NN algorithm computes the distance for each training and testing sample in K-DD data sets to find exact nearest neighbors . Furthermore, it is very effective to deal with big data. However, we have some sequence of data (x_1, y_1), . . ., (x_n, y_n), to be access $x_i \in \boldsymbol{R}^d$ and binary classification $y_i \in \{0, 1\}$; the Euclidean distance is used as the distance metric to measure the similarity between two points [7].

$$D = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \quad (1)$$

$$d^2(x_i - x_j) = \|x_i - x_j\|^2 = \sum_{k=1}^{d} (x_{ik} - x_{jk})^2$$

$$(x_i, x_j) \in R^d, x_i = (x_{i1}, x_{i2}, \dots, x_{id}).$$

Classify sub_testdata by using K-NN classifier.

3.2. BA Algorithm

Bees Algorithm(BA) is very effective algorithm to find the good food location and mimics the behavior of honeybees. Furthermore, can be used for combination and functional optimization. The process of optimization is carried ant in order to find ant the combination and the function of the feature selected in the intrusion detection system(IDS) [9]. The feature selection is the step which follows the optimization, combination and then selection of the items and features from the big data system. This is what is known as the intrusion detection system(IDS) [10].

- Initialize population with random solutions.

- Appraisal fitness of the population.

- Choose elite bees.

- Choose sites for neighborhood search.

- Enlist bees for choose sites.

- Choose the appropriate bees from every spot.

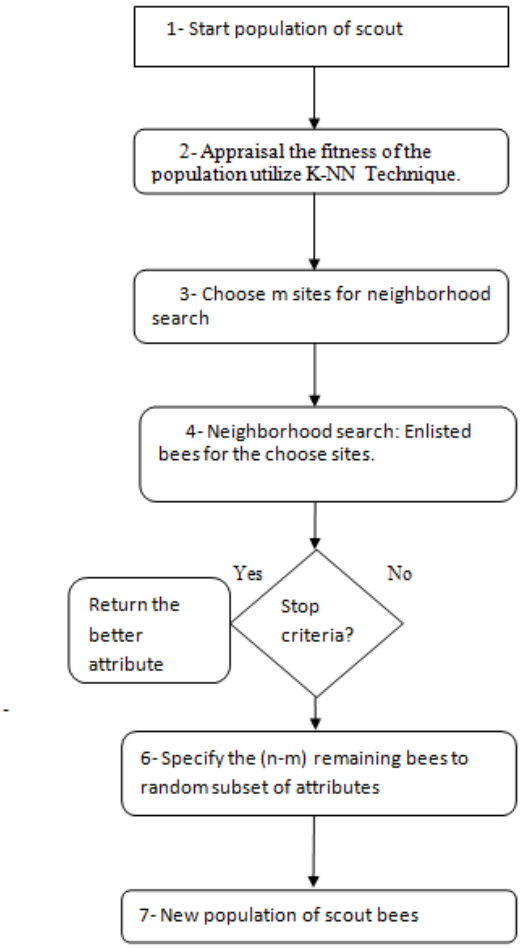
- End while .

Figure 1. Pseudocode for the BA

3.3. Neighborhood search

The purpose of Neighborhood searches to select the number of features randomly. There are 41 features in KDD datasets. Selected population Number for creating sub-datasets with sub-selected features for training and testing . So, recruit bee of lowest fitness into other bee using crossover method. In KDD cup99, there are several steps including preprocessing of data, reduction of features data, classification and evaluation. In KDD, there are 4 main attack types including 'DOS', 'Probing', 'R2L' and 'U2R', 22 attack types in detail. So we used one method to reduce the dimensionality of data while preserving information of original data. That is Genetic Algorithm(GA) in KDD . Before that, we had to convert the KDD cup99 data into numeric data with its elements is 0 or 1 and this step was implemented by preprocessing .The KDD data has 4 string features and then we counted the number of the each property of string features and set number according to big or small. And used k-NN(k-nearest neighbors) method for classifying KDD 99cup.

The generic model proposed of K-NN-BA



There are three features of KDD dataset, this is (I) protocol type (II) service attribute and (III) flag attributes. These three attributes are useful for classification then mapped. Data preprocessing has the second step after the initial step of data mapping which is of identification of state. This step shows that there are five categories in the attacks on big data

IV. Experiments and results

The selected the ration training data against whole KDD as about 70% and testing as 30% in the performance. The Probe, Dos, U2R, R2L is selected in proportion to total numbers of own cases uniformly in 30%. The population number is 20, extracted features number is 8, crossover number is 2 and constant between 0 and 1 to decide fitness in BA algorithm was 0.8. The main reason is to extract best features of data which are represented well property of Groups using GA(genetic algorithm). The total records used for training and testing about 494201 connection records.

4.1. Appraisal standard

Applied three measures for evaluate results in the experiments as follow: accuracy rate (AR), detection rate (DR), and false Positive rate (FPR). This appraisal standard is denoted as following equations.

The detection rate is defined as the proportion of aggregate numbers of attacks correctly classified which using K-NN to the aggregate numbers of attacks in the KDD99 datasets.

$$DR = \frac{Number\ of\ attacks\ correctly\ classified\ as\ attacks}{Total\ number\ of\ attacks\ in\ kdd99\ dataset} \qquad (2)$$

The false positive rate over a network is defined as the proportion of total numbers of normal events which using K-NN to the aggregate numbers of normal datasets.

$$FPR = \frac{\textit{Number of normal events classified as attack}}{\textit{Total number of normal events in the kdd99 dataset}} \quad (3)$$

The accuracy rate is defined as the proportion numbers of classified instances which using K-NN to aggregate numbers of instances in the KDD99 datasets.

$$ACR = \frac{\textit{Number of correctly classified instances}}{\textit{Total number of instances in the kdd99 datasets}} \quad (4)$$

4.2. Fitness Function

The population which using K-NN algorithm so the fitness = alpha*DR + (1-alpha)*FAR where alpha is any number between 0 and 1 according to an importance of detection or false positive rate. DR = aggregate numbers of attacks correctly classified which using K-NN to the aggregate numbers of attacks in the KDD99 datasets whereas FAR = aggregate numbers of classified instances which using K-NN to aggregate numbers of instances in the KDD99 datasets. best fitness = 0; bestId = 0; repeat = 0; subtrain data = zeros(rows, reduction No, population No); subtest data = zeros(trows,reductionNo,populationNo);classify=zeros(trows,populationNo).

4.3. Results

The proposed K-NN-BA approaches is implemented used MATLAB R2015a-64 bit installed on windows 7 Ultimate with the core i5 processor and 12 GB RAM . Table 1 shows the obtained results of the proposed model in three measures DR, FPR, and AR .

Table 1. Results of the proposed K-NN-BA

Name	***Value***	***Description of value***
Data	494021	Numerical data of which converted
Classify	148207	Label classified of testing
Num1	346112	The whole of data used for training
Attack	277857	Total number of attacks in training data
Normal	68255	Total number of normal in training data
Num2	147909	Total number of testing data
Tattack	118886	Total of attacks number in testing data
Tnormal	29023	Total of normal number in training data
Subfeature	5,24,30,6,23,25,34,26	Selected 8 best features of data
Subtraindata	346112	Sub-training data extracted using sub-features
Subtestdata	147909	Sub-testing data extracted using sub-features
DR	99. 88	Detection Rate
AR	99.52	Accuracy Rate
FBR	0.48	False Positive Rate
Time	0.874 (second)	Calculate time of DR, AR, FBR

The obtained results with significant features gave higher detection and accuracy rate whereas reducing false positive rate.

Table 2. Result of the selected features by the k-NN-BA approach

Number of attribute	**Attributes name**	**Data type**
1	src_bytes	Continuous
2	dst_bytes	Continuous
3	Count	Continuous
4	srv_count`	Continuous
5	serror_rate	Continuous
6	srv_serror_rate	Continuous
7	diff_srv_rate	Continuous
8	dst_host_same_srv_rate	Continuous

Table 3. Most significant of attributes.

Attribute	Description
src_bytes	Bytes transferred from source to destination
dst_bytes	Bytes transferred from destination to source
count	Same- host connections
srv_count	Same- service connection
serror_rate	Same- host connections but have ``SYN" errors
srv_serror_rate	Same- service connection but have``SYN" errors
diff_srv_rate	Connection to different services

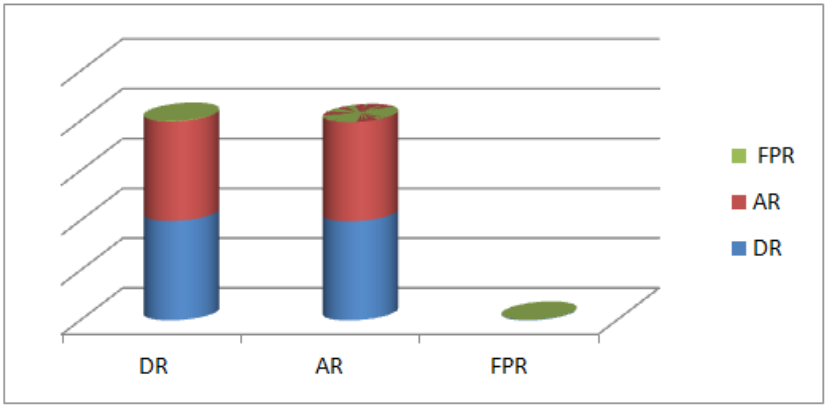


Figure 3: Results of DR, AR, and FAR

V. Conclusion and Future work

The present paper shows how two algorithms are useful in the process of big data intrusion detection system(IDS). These two algorithms are bees algorithm(BA) and k-nearest neighbor (K-NN). The data has been reduced for extracting features by Bees Algorithm (AB) whereas applied K-nearest neighbors as classification (KNN). So, the KDD99 datasets applied in the experiments with significant features. for this reason, the proposed model gave a high increasing of detection and accuracy rate. On the other hand, The model gives decrease of false positive rate.

Reference

[1]. A.Suebsing, N. Hiransakolwong "Euclidean-based Feature Selection for Network Intrusion Detection" 2009 International Conference on Machine Learning and Computing IPCSIT vol.3 (2011) IACSIT Press, Singapore.
[2]. R.-Ching Chen and K.Fan Cheng **"**Using Rough Set and Support VectorMachine for Network Intrusion Detection" International Journal of NetworkSecurity & Its Applications (IJNSA),Vol 1, No 1, April 2009.
[3]. K. Rufai, R.Chandren **"**Improving Bee Algorithm Based Feature Selection in Intrusion Detection System Using Membrane Computing" Journal of Networks, vol. 9, no. 3, March 2014.
[4]. A. Sabry and E. Zeynep " A new feature selection model based on ID3 and bees algorithm for intrusiondetection system"Turkish Journal of Electrical Engineering & Computer Sciences Turk J Elec Eng & Comp Sci(2015) .
[5]. A. Amin, M. Bin Ibne Reaz "A novel SVM-kNN-PSO ensemble method for intrusion detection system" Department of Electrical, Electronic & Systems Engineering, Faculty of Engineering & Built Environment, National University of Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia 2016.
[6]. Zhenyun.D,Xiaoshu.Z"Efficient kNN classification algorithm for big data "Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China, 2016 Elsevier.
[7]. Jingwei H, Zbigniew K. "Knowledge Discovery from Big Data for Intrusion Detection Using LDA"
[8]. R. Mall, V. Jumutc "Representative Subsets For Big Data Learning using k-NN Graphs" European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)
[9]. Osama.A. Zulaiha. A. "Bees Algorithm for feature selection in Network Anomaly detection" Journal of Applied Sciences Research 2012.
[10]. Valery T, Andreas. L. "Collective Decision-Making in Honey Bee Foraging Dynamics" October 13, 2005.