ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

BEHAVIOR BASED INVADING USING BAGGING

P.Akshaya,
PG Student,
Dept of Computer Science,
Thiagarajar college of engineering,
Madurai.

Dr.K.Sundarakantham, Asst prof,
Dept of Computer Science,
Thiagarajar college of engineering,
Madurai.

Abstract

Intrusion detection system is the system of preventing the intruders to getting inside either knowingly or unknowingly. Intrusion detection systems act as a barrier for various attacks that are made in the network. The performance of the intrusion detection system is based on the classification accuracy and false positive rates. It is the challenge in today's world of network to design an intrusion detection system with the highest detection rate with low false positives. With a thorough survey on the intrusion detection system and the machine learning algorithms, in this paper bagging ensemble method of machine learning is concluded to be used. Decision trees make good candidates for combining because they are structurally unstable classifiers, and produce diversity in classifier decision boundaries. Small perturbations in training data set can result in very different model structures and splits. The model is based on the misuse detection and uses c4.5 classifier. Hence, the proposed model will be accurate and simple owing to the usage of decision tree classifier.

Keywords: Ensemble, Decision tree, Intrusion detection.

1 Introduction

In recent years, there has been an explosion discussing how to combine models or model predictions and the reduction in model error that results. By combining predictions, more robust and accurate models nearly always improve without the need for the high-degree of fine tuning required for single-model solutions. Typically, the models for the combination process are drawn from the same algorithm family (decision trees), though this need not be the case. Bagging combines outputs from decision tree models generated from bootstrap samples (with replacement) of a training data set. Models are combined by simple voting of

the individual model output predictions. Bagging takes the base classifier and training set as its input. Decision trees use "stair-step" decision boundaries produced by if-then rules. Combining decision tree models has the positive benefit of smoothing these decision regions, thus improving the robustness of the classifier. Boosting and bagging are almost used exclusively with decision trees in the research and software products today. Since, the supervised algorithms are commonly described as performing the task of searching through a hypothesis space to find a suitable hypothesis that will make good predictions with a particular problem. Even if the hypothesis space contains hypotheses that are very well-suited for a particular problem, it may be very difficult to find a good one. Ensembles combine multiple hypotheses to form a better hypothesis. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner. The broader term of multiple classifier systems also covers hybridization of hypotheses that are not induced by the same base learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. Fast algorithms such as decision trees are commonly used with ensembles. Bootstrap aggregating, often abbreviated as bagging involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set. The proposed model uses the bagging ensemble method of machine learning to design an intrusion detection system wherein the decision trees are used as a base classifier.

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

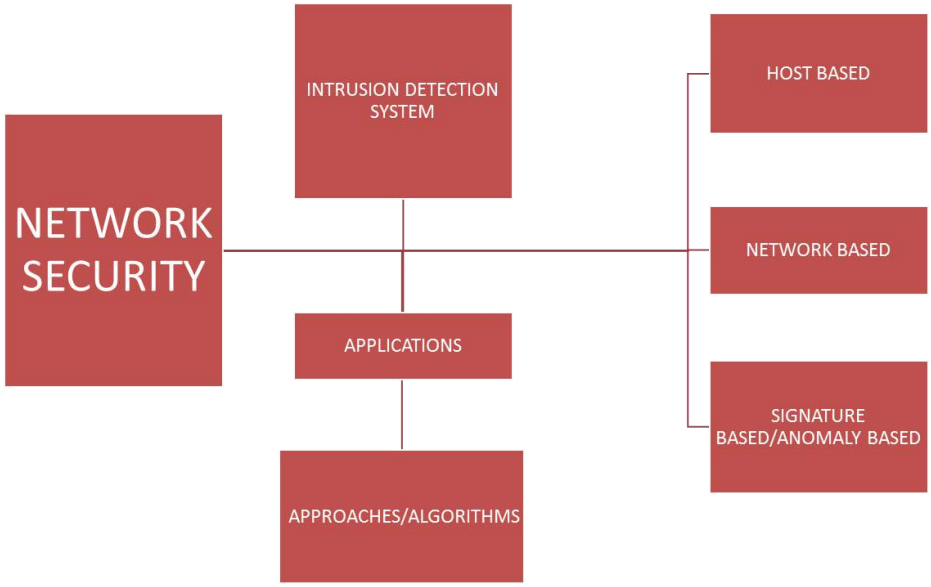


Fig 1. Classifications of intrusion detection systems

This paper is composed of 7 sections. Section 1 deals with the introduction. Section 2 is about the existing model and proposed model of intrusion detection system. Section 3 deals about system architecture. Section 4 explains the bagging ensemble. Section 5 deals with the results and Section 6 is about Conclusion.

2 Intrusion Detection Systems

2.1 Existing System

Genetic algorithm weight extraction algorithm is used in the optimization of weights between the neurons of ANN to identify the intrusion in an efficient manner. In particular, several Neural Networks based approaches were employed for Intrusion Detection. Several Genetic Algorithms (GAs) has been used for detecting Intrusions of different kinds in different scenarios [6][7] [8] [9]. GAs used to select required features and to determine the optimal and minimal parameters of some core functions in which different AI methods were used to derive acquisition of rules [10] [11] [12]. In [13], authors presented an implementation of GA based approach to Network Intrusion Detection using GA and showed software implementation. The approach derived a set of classification rules and utilizes a support-confidence framework to judge fitness function. In [14], authors designed a GA based performance evaluation algorithm for network

intrusion detection. The approach uses information theory for filtering the traffic data. In [15], authors used the BP network with GAs for enhancement of BP; they used some types of attack with some features of KDD data. A back-propagation Neural Network was used [16], authors used all features of KDD data, the classification rate for experiment result for normal traffic was 100%, known attacks were 80%, and for unknown attacks were 60%.

Disadvantages:

- This system will not give accurate results.
- It cannot predict the correct user activity.
- Compared with the proposed one the process level is low.

2.2 Proposed system

Genetic Algorithm is chosen to make this intrusion detection system. This section gives an overview of the algorithm and the system. Genetic Algorithm (GA) is a programming technique that mimics biological evolution as a problem-solving strategy [17]. It is based on Darwinian's principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness [7]. GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators [7]. The process usually begins with randomly generated population of chromosomes, which represent all possible solution of a problem that are considered candidate solutions. From each chromosome different positions are encoded as bits, characters or numbers. These positions could be referred to as genes. An evaluation function is used to calculate the goodness of each chromosome according to the desired solution; this function is known as "Fitness Function". During the process of evaluation "Crossover" is used to simulate natural reproduction and "Mutation" is used to mutation of species [7]. For survival and combination the selection of chromosomes is biased towards the fittest chromosomes. When we use GA for solving various problems three factors will have vital impact on the effectiveness of the algorithm and also of the applications [18]. They are: i) The fitness function; ii) The representation of individuals iii) The GA parameters. The determination of these factors often depends on applications and/or implementation. Also all the three steps of generating new population from old population are depicted. The process of generating new population from old population includes selection, crossover, and mutation. If new population is not feasible then quit, otherwise again repeat the

generation process. This system can be divided into two main phases: the pre calculation phase and the detection phase. Following are the major steps in pre calculation phase, where a set of chromosome is created using training data. This chromosome set will be used in the next phase for the purpose of comparison. C4.5 is an algorithm used to generate a decision trees and an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used as a statistical classifier, since it is best in classification. It uses the concept of information entropy and builds the decision trees similar to the ID3 algorithm. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists. Neural Networks (NNs) have attracted more attention compared to other techniques. That is mainly due to the strong discrimination and generalization abilities of Neural Networks that utilized for classification purposes [19]. Artificial Neural Network is a system simulation of the neurons in the human brain [20]. It is composed of a large number of highly interconnected processing elements (neurons) working with each other to solve specific problems. Each processing element is basically a summing element followed by an active function. The output of each neuron (after applying the weight parameter associated with the connection) is fed as the input to all of the neurons in the next layer. The learning process is essentially an optimization process in which the parameters of the best set of connection coefficients (weights) for solving a problem are found [21]. An increasing amount of research in the last few years has investigated the application of Neural Networks to intrusion detection. If properly designed and implemented, Neural Networks have the potential to address many of the problems encountered by rule-based approaches. Neural Networks were specifically proposed to learn the typical characteristics of system's users and identify statistically significant variations from their established behaviour. In order to apply this approach to Intrusion Detection, I would have to introduce data representing attacks and non-attacks to the Neural Network to adjust automatically coefficients of this Network during the training phase. In other words, it will be necessary to collect data representing normal and abnormal behaviour and train the Neural Network on those data. After training is accomplished, a certain number of performance tests with real network traffic and attacks should be conducted [22]. Instead of processing program instruction sequentially, Neural Network based models on simultaneously

explorer several hypotheses make the use of several computational interconnected elements (neurons); this parallel processing may imply time savings in malicious traffic analysis . KDD 99 data set are used as the input vectors for training and validation of the tested neural network. It was created based on the DARPA intrusion detection evaluation program. MIT Lincoln Lab that participates in this program has set up simulation of typical LAN network in order to acquire raw TCP dump data. They simulated LAN operated as a normal environment, which was infected by various types of attacks. The raw data set was processed into connection records. For each connection, 41 various features were extracted. Each connection was labelled as normal or under specific type of attack. There are 39 attacker types that could be classified into four main categories of attacks: DOS (Denial of Service): an attacker tries to prevent legitimate users from using a service. E.g. TCP SYN Flood, Smurf (391458 record). Probe: an attacker tries to find information about the target host. For example: scanning victims in order to get Knowledge about available services, using Operating System (4107 record). U2R (User to Root): an attacker has local account on victim's host and tries to gain the root privileges (52 records). R2L (Remote to Local): an attacker does not have local account on the victim host and try to obtain it (1124 records).

Advantages:

- It is used to improve Accuracy classification performance.
- It is used to predict movement accuracy.
- It is sufficient.
- It will be make new population.
- It produces low false positives.
- The model building time is low.
- It takes the advantages of the ensemble machine learning method where bagging is the added credit the system.

3 System Architecture

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

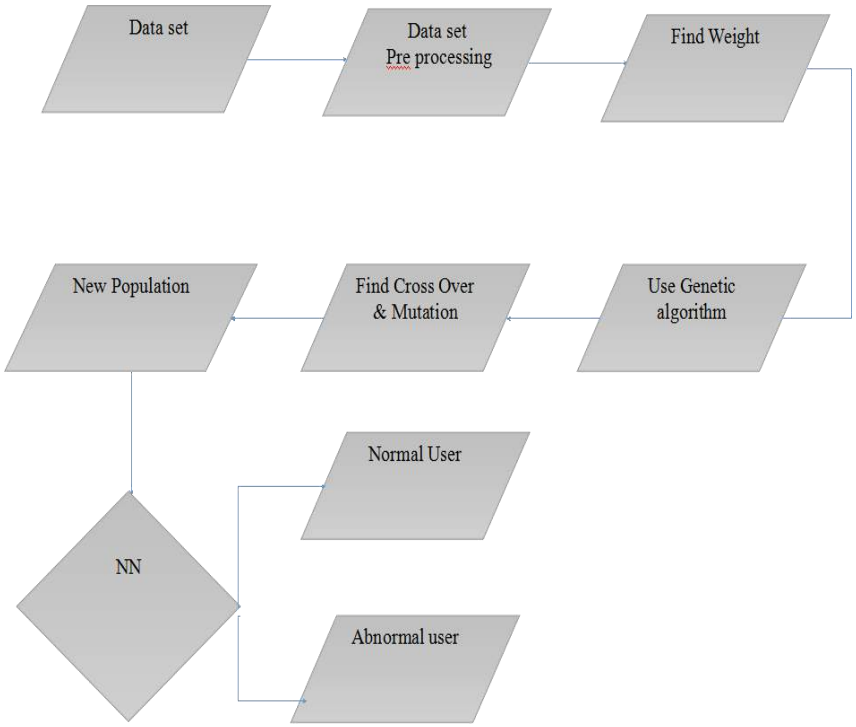


Fig 2. Process flow of the behaviour based bagging IDS

4 Bagging Ensembles

4.1 Features Selection

The available data sets to build the intrusion detection system are the DARPA 1998, NSL –KDD 99, KDD. The training data set that is being used in this work is the NSL -KDD 99.There are totally 41 features that have been suggested by the NSL-KDD. The usage of all the features will result in less accuracy and more model building time. Hence, only 15 features are decided to be used to get the highest accuracy and low model building time. To prune the features we do pre processing since the feature selection has the advantages such as reducing dimension, boosting generalization capability, accelerate learning and model interpretation. The feature selection is done using the genetic algorithm.

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

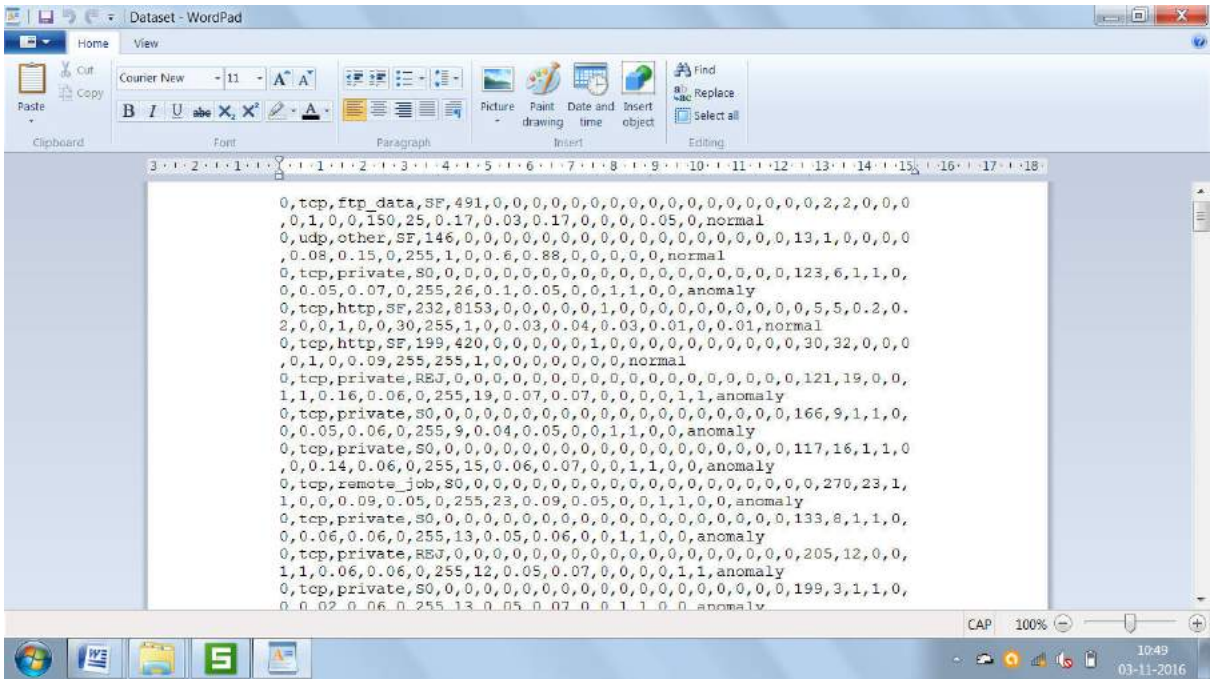


Fig 3. KDD 1998 Data set in form of csv

4.2 Pre processing

This process is first convert comma separated values(csv) file to text. Then it will be remove unwanted data from text file. The pre processing involves the technique called categorical encoding. By which the non numeric values will be converted into numeric by arbitrarily assigning values to each category. This pruning techniques helps in time conception and memory conception . The processed data set is then loaded into the database. Then the data is normalized by using the values of the attribute. The normalization is the act of enclosing the values of the attributed to a specific range to minimize the complexity involved in dealing with data spread over an absolute range and type of values. The normalization is carried out using the below formulae,

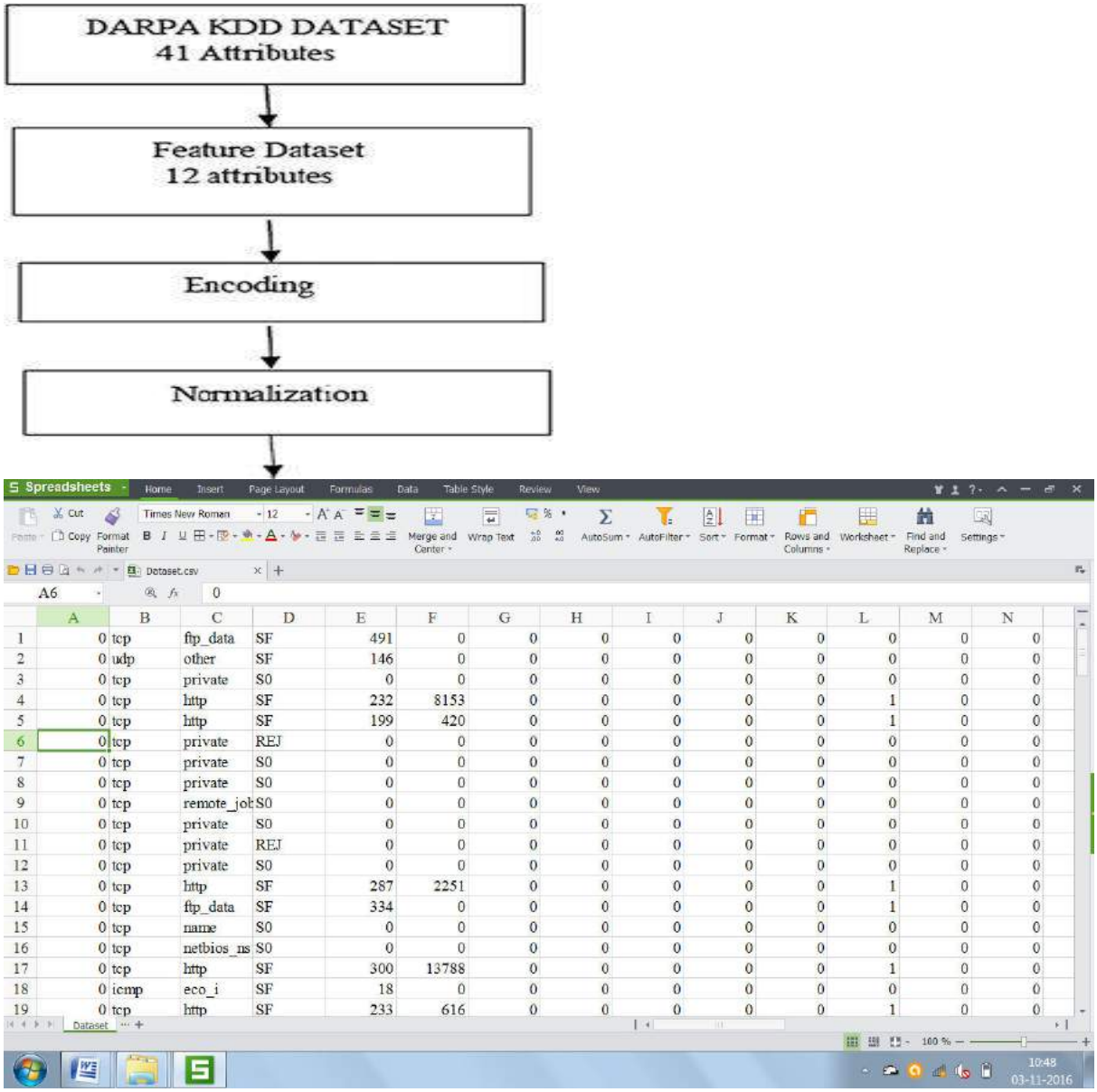
V`=((V-min A)/(max A-min A)) * ((new max A-new min A) + new min A)

Where,
V` is the value of the attribute after normalization.
V is the value of the attribute before normalization.

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

min A is the minimum value of the attribute observed.
max A is the maximum value of the attribute observed.
new min A is the minimum value of the attribute at new range.
new max A is the maximum value of the attribute at new range.
A is the attribute.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0	0
2	0	udp	other	SF	146	0	0	0	0	0	0	0	0	0
3	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0
4	0	tcp	http	SF	232	8153	0	0	0	0	0	1	0	0
5	0	tcp	http	SF	199	420	0	0	0	0	0	1	0	0
6	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0
7	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0
8	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0
9	0	tcp	remote_job	S0	0	0	0	0	0	0	0	0	0	0
10	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0
11	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0
12	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0
13	0	tcp	http	SF	287	2251	0	0	0	0	0	1	0	0
14	0	tcp	ftp_data	SF	334	0	0	0	0	0	0	1	0	0
15	0	tcp	name	S0	0	0	0	0	0	0	0	0	0	0
16	0	tcp	netbios_ns	S0	0	0	0	0	0	0	0	0	0	0
17	0	tcp	http	SF	300	13788	0	0	0	0	0	1	0	0
18	0	icmp	eco_i	SF	18	0	0	0	0	0	0	0	0	0
19	0	tcp	http	SF	233	616	0	0	0	0	0	1	0	0

Fig 4. Data set after pre processing (ie) Featured data set.

4.3 Genetic algorithm:

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

Genetic algorithm uses global optimization searching and simulates the behaviour of evolution process in nature. It maps the searching space into genetic space. In genetic algorithm, a chromosome (also sometimes called a genotype) is a set of parameters which define a proposed solution to the problem. The chromosome is often represented as a simple string; although a wide variety of other data structures are also used. The chromosomes are estimated according to fitness function.

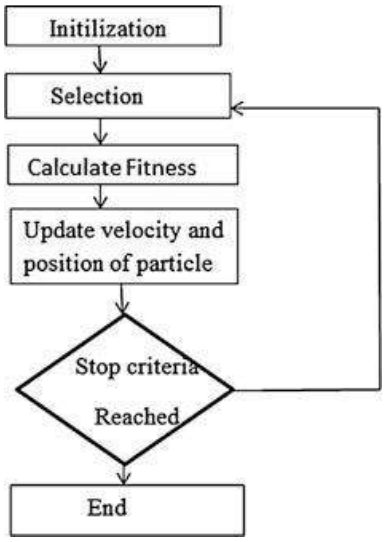


Fig 5. Flow of the Genetic Algorithm

4.4Neural Networks:

A typical neural network has anything from a few dozen to hundreds, thousands, or even millions of artificial neurons called units arranged in a series of layers, each of which connects to the layers on either side. Some of them, known as input units, are designed to receive various forms of information from the outside world that the network will attempt to learn about, recognize, or otherwise process. Other units sit on the opposite side of the network and signal how it responds to the information it's learned; those are known as output units.

4.5 Performance Analysis:

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

An attack pattern, which may be an attacker-specific pattern or a pattern commonly used by attackers, can be identified in the same method. Similarly, an attack pattern that an attacker frequently submits but others have seldom or never submitted will be considered as one of the attacker's representative attack patterns and will obtain a high similarity weight. Hence, signatures collected in an attacker profile can be classified into common signatures and attacker-specific signatures. The latter can be used to identify who the possible attackers are when a protected system is attacked by attacker-specific signatures.

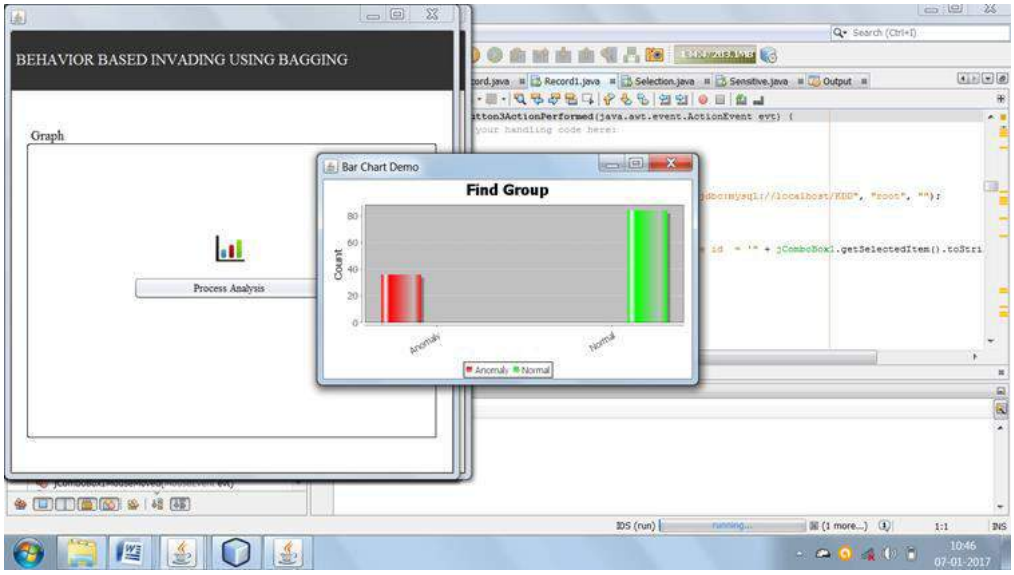


Fig 6. Result of decision tree classification algorithm in form of bar chart

5 Results

	Positive	Negative
True	32.0	40.0
False	37.0	25.0

Table 1. Prediction

ACCURACY ANALYSIS	VALUE

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Special Issue.24, March 2017

Sensitivity	21.0
Specificity	0.181176
True Prediction	0.44
False Prediction	**0.2566**

Table 2. Analysis

6 Conclusions

In this work, the proposed an IDS system is used to detect various types of attacks with high accuracy and low false positives. The proposed system is providing both type of functionality in one system which is improving overall efficiency of the existing IDS. In future we will we work on network layers protocol and try to find attack on network layers that mean in this layers what type of attack will perform and how we can protect from or prevent them.

The proposed IDS system to detect various types of layers attacks like application layer, transport layer and abnormal packets in N/W IDS and additionally in Host IDS UN-authorize accessing and login/logout failed. The proposed system is providing both type of functionality in one system which is improving overall efficiency of the existing IDS. In future we will we work on network layers protocol and try to find attack on network layers that mean in this layers what type of attack will perform and how we can protect from or prevent them. The score-based multi-cyclic SR algorithm outperformed the multi-cyclic CUSUM procedure. Lastly, as a possible improvement of any change point detection-based anomaly detector, we proposed to complement the latter with a signature-based spectral ID. This approach will allow filtering false alarms reducing the false alarm rate to a minimum and simultaneously guaranteeing prompt detection of real attacks.

7 References

- [1] Data Stream Based IDS for Advanced Metering infrastructure in Smart Grid. Mustafa Amir Fasil,John R.Williams. IEEE Sytems Journal,vol.9,No.1,March 2015.

- [2] Efficient Computer Network Anomaly detection by change point detection methods: Alexander G. Tartakovsky, *Senior Member, IEEE*, Aleksey S.Polunchenko, and Grigory SokolovIEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 7, NO. 1, FEBRUARY 2013.
- [3] A semantic approach to host based intrusion detection system using contiguous and discontiguous system call patterns gideon creech, student member, IEEE TRANSACTIONS ON COMPUTERS,VOL.63,NO.4,APRIL 2014.
- [4] Network intrusion detection system embedded on smart sensor Fransico Macia, Iren lorenzo-Fonesca. IEEE Transactions On industrial Electronics,Vol.58,No.3,March 2011.
- [5] An Effective and feasible traceback scheme in mobile internet environment Federico Maggi,Matteo Matteucci IEEE systems journal .Vol.18,No.11,NOV 2014.
- [6] Detecting intrusions through system call sequence and argument analysis Federico Maggi, Student Member, IEEE, Matteo Matteucci, Member, IEEE, and Stefano Zanero, Member, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 7, NO. 4, OCTOBER-DECEMBER 2010.
- [7] Bayesian based intrusion detection system Hesham Altwaijry, Saeed Algarny Computer Engineering Department, King Saud niversity, P.O. Box 51178, Riyadh 11543, Saudi Arabia Received 14 December 2010; accepted 24 April 2011.
- [8] IDS using partial decision tree classifier Procedia Computer Science 49 (2015) 92 – 98.
- [9] Effective approach toward intrusion detection system using data mining Techniques G.V. Nadiammai, M. Hemalatha Egyptian Informatics Journal (2014) 15, 37–50.

- [10] J., Muna. M. and Mehrotra M., "Intrusion Detection Systems : A design perspective", Proceeding of 2rd International Conference On Data Management, IMT Ghaziabad, India.,2009,265-372.
- [11] M. Panda, and M. Patra, "Building an efficient network intrusion detection model using Self Organizing Maps", Proceeding of world academy of science, engineering and technology, 38, 2009, 22-29.
- [12] Zhang, K., Cao, H.-x., Yan, H.: Application of support vector machines on network abnormal intrusion detection. Application Research of Computers 5, 98–100 (2006).
- [13] Bosin, A., Dessì, N., Pes, B.: Intelligent Bayesian Classifiers in Network Intrusion Detection. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533,p. 445–447. Springer, Heidelberg (2005).
- [14] Rawat, S., Sastry, C.S.: Network Intrusion Detection Using Wavelet Analysis. In: Das, G., Gulati, V.P. (eds.) CIT 2004. LNCS, vol. 3356, pp. 224–232. Springer, Heidelberg (2004).
- [15] Guan, J., Liu, D.-x., Wang, T.: Applications of Fuzzy Data Mining Methods for Intrusion Detection Systems. In: Laganá, A., Gavrilova, M.L., Kumar, V., Mun, Y., Tan, C.J.K., Gervasi, O. (eds.) ICCSA 2004. LNCS, vol. 3045, pp. 706–714. Springer, Heidelberg(2004).
- [16] Kim, D.S., Park, J.S.: Network-based intrusion detection with support vector machines. In: Kahng, H.-K. (ed.) ICOIN 2003. LNCS, vol. 2662, pp. 747–756. Springer, Heidelberg(2003).
- [17] Rao, X., Dong, C.-x., Yang, S.-q.: An intrusion detection system based on support vector machine. Journal of Software 4, 798–803 (2003).
- [18] llgun, K.: USTAT: a real-time intrusion detection system for UNIX. In: Proceedings of the 1993 Computer Society Symposium on Research in Security and Privacy, pp. 16–29 (1994).
- [19]KDDCup99Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
- [20] Debar, H., Dorizzi, B.: An application of a recurrent network to an Intrusion Detection System. In: Proceedings of IJCNN, pp. 78–83 (1993).

- [21] B. Mykerjee, L. Heberlein T., and K. Levitt N., "Network Intrusion Detection", IEEE Networks, 8(3), 1992, 14-26.