Using Gath-Geva Fuzzy clustering algorithm and fuzzy Back Propagation Neural Network for intrusion detection and classification of Attacks

Assist. Prof. Dr. Baydaa Ibraheem Khaleel

Dept. of Computer Science, College of Computer Science & Mathematics, Mosul University, IRAQ

ABSTRACT

Classification or (cluster analysis) has been widely used in data analysis and pattern recognition. And along with the development and growth of the internet network, there is an increasing needed to protect computer and network from attacks and unauthorized access. Such that network intrusion classification and detection systems to prevent unlawful accesses. We have take the advantage of classification abilities of fuzzy clustering algorithms to recognize intrusion(attack) and also detect attacks. There are several algorithms for classifying and clustering large data sets or streaming data sets, Their aims to organize a collection of data items into groups. These such items are more similar to each other within group (class), and difference than they are in the other classes. Gath-Geva fuzzy clustering algorithm, and combined Gath-Geve with back propagation neural network to produce Fuzzy Back Propagation network (FBPN) algorithm were applied using NSL-KDD data set to classify this data set into 5 classes according to the type of attacks (Normal, DoS, Probe, U2R, R2L). We compute the classification rate, detection rate on this data set. Finally we make comparisons between results obtained after applying the algorithms on this data set.

Keywords: Gath-Geve (GG) fuzzy clustering algorithm, fuzzy Back propagation neural network, Intrusion detection, NSL-KDD data set.

HOW TO CITE THIS ARTICLE

Assist. Prof. Dr. Baydaa Ibraheem Khaleel, “Using Gath-Geva Fuzzy clustering algorithm and fuzzy Back Propagation Neural Network for intrusion detection and classification of Attacks”, International Journal of Enhanced Research in Science, Technology & Engineering, ISSN: 2319-7463, Vol. 8 Issue 7, July-2019.

1- INTRODUCTION

An intrusion detection system (IDS) is a component of the information security framework. Its main goal is to differentiate between normal activities of the system and behavior that can be classified as suspicious or intrusive. The goal of intrusion detection is to build a system which would automatically scan network activity and detect such intrusion attacks [1][2]. Once an attack is detected, the system administrator can be informed who can take appropriate action to deal with the intrusion [2]. The number of intrusion into computer systems is growing because new automated intrusion tools appearing every day, and these tools and different system vulnerability information are easily available on the web [3]. Using an intrusion detection system (IDS) is one way of dealing with suspicious activities within a network[4]. Intrusion detection, an important component of information security technology, helps in discovering, determining, and identifying unauthorized use, and destruction of information and information systems[5][6]. The goals of intrusion detection are detect as many type of attacks as possible, including those by attackers and those by insiders. Also detect as accurately as possible thereby minimizing the number of false alarms, and also detect the attacks in the shortest possible time[7]. Intrusion detection techniques can be categorized into misuse detection and anomaly detection[1] .

- misuse detection uses the patterns of well-known attacks or vulnerable spots in the system to identify intrusions [8]. Misuse detection is based on the knowledge of system vulnerabilities and known attack patterns. Misuse detection is concerned with finding intruders who are attempting to break into a system by exploiting some known vulnerability, ideally, a system security administrator should be a were of all the known vulnerabilities and eliminate them [9].
- Anomaly detection attempts to determine whether can be flagged as intrusions.

There are three types of intrusion detection systems: Host-based Intrusion Detection System (HIDS), Network-based Intrusion Detection System (NIDS), and combination of both types (Hybrid Intrusion Detection System) [2][8].

2- PREVIOUS WORK

In particular several classification or clustering algorithms and artificial intelligence techniques were used for intrusion detection and classification. In 2016 Urvashi M., and Anurag J.[1] obtained Detection rate 96 % of three machine learning algorithm J48, J48 Graft and Random Forrest were used. In 2015 Liu X., Tian J.,**[**5 **]** use traditional K-means algorithm Detection rate 90.0% , and improved dot density clustering algorithm MSTK-means algorithm Detection rate 98.00%. Siddiqui[10] used parallel back propagation neural network and pararllel fuzzy ARTMAP, the detection rate result for parallel BP in training stage is 98.36 and the detection rate in testing stage is 81.73 and false alarm is 1.28. Detection rate for parallel fuzzy ARTMAP in training stage is 80.14 and in testing state detection rate is 80.52 and false alarm is 19.48. Al-Sharafat and SH. Naoum [11] used Steady State Genetic Algorithm Based Machine Learning SSGBML, and used kdd 99 dataset, the detection rate for this approach is 97.45 in training state.

3- NSL-KDD DATASET

NSL is a new version of KDDcup99 and has some advantages over KDD cup 99, which is also contains 41 features and labeled as either normal or attack as the same in KDD cup 99, the NSL-KDD data set has the following advantages over the original KDD data set[12][13]:

- It does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records.
- There is no duplicate records in the proposed test sets; therefore, the performance of the learners are not biased by the methods which have better detection rates on the frequent records.
- The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. As a result, the classification rates of distinct machine learning methods vary in a wider range, which makes it more efficient to have an accurate evaluation of different learning techniques.
- The number of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

The total number of connection records in training data set of NSL is “kddTrain+.TXT” file that contain (125973)records, and the total number of connection records in testing data set is “kddTest+.TXT” file that contain on (22544) records[9].

4- NEURAL NETWORKS

A neural network represents a highly parallelized dynamic system with a directed graph topology that can receive the output information by means of reaction of its state on the input nodes. The ensembles of interconnected artificial neurons generally organized into layers of fields include neural networks. The behavior of such ensembles varies greatly with changes in architectures as well as neuron signal functions [13]. Well-trained neural networks represent a knowledge base in which knowledge is distributed in the form of weighted interconnections where a learning algorithm is used to modify the knowledge base from a set of given representative cases. Neural networks might be better suited for unstructured problems pertaining to complex relationships among variables rather than problem domains requiring value-based human reasoning through complex issues. [14].

5- PREPROCESSING DATA

From NSL-KDD intrusion detection dataset, 41 features were derived to summarize each connection information. In order to train an architecture, several data enumeration and normalization operations were necessary. As a first approach, symbolic variables in the dataset were enumerated and all variables were normalized. Thus, each instance of a symbolic feature was first mapped to sequential integer values[15]. This dataset consist of symbolic and numeric values, all symbolic values were transformed into numeric values such as three types of protocols (tcp, udp, icmp) and 71 type of services in NSL-KDD and 11 types of flag, each one take value from [1..N], and the standard [0..1] normalization[16] was used for this research according equation(1):

$$X = \frac{x - \min}{\max - \min} \qquad (1)$$

Where, X is the numerical value, $\min$ is the minimum value for the attribute that x belongs to, and $\max$ is the maximum value for the attribute that x belongs to.

6- PERFORMANCE MEASURES

The indicators were used to measure the accuracy of the methods: classification rate and detection rate. The classification rates as shown in equation (2). While the detection rate shows the percentage of true intrusions that have been successfully detected as shown in equation(3)[17][18].

$$classifica\ \ tion\ \ rate\ \ (CR) = \frac{number\ \ of\ \ samples\ \ classified\ \ correctly}{number\ \ of\ \ samples\ \ used\ \ for\ \ training} \times 100 \qquad (2)$$

$$Detection\ \ \_\ rate\ \ (DR) = \frac{number\ \ of\ \ correctly\ \ \det\ ected\ \ samples}{total\ \ number\ \ of\ \ samples} \times 100 \qquad (3)$$

7- GATH-GEVA FUZZY CLUSTERING ALGORITHM

Gath-Geve (GG) can be used to detect ellipsoidal clusters with varying size[19]. G-G fuzzy clustering algorithm takes the size and density of clusters for classification[20]. The objective function based on the minimization of the sum of weighted squared distances between the data points and cluster centers is described in the following[21][19]:

$$J\ (\mathrm{Z}, U, \mathrm{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{\mathrm{ik}})^{m}\ D_{ik}^{2} \qquad (4)$$

Where Z is the set of data, $U = [\mu_{\mathrm{ik}}]$ is the fuzzy partition matrix, $V = [V_1, V_2, ..., V_c]^T$ is the set of centers of the clusters, c is the number of clusters, N is the number of the data, m is the fuzzy coefficient, μ_{ik} is the membership degree between the i-th cluster and k-th data, which satisfies conditions:

$$\mu_{ik} \in [0,1]; \sum_{i=1}^{c} \mu_{ik} = 1 \qquad (5)$$

The minimum of (U, V) is calculated as follows:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (D_{ik} / D_{jk})^{2/(m-1)}} \qquad (6)$$

$$v_i = \frac{\sum_{\mathrm{k}=1}^{\mathrm{N}} (\mu_{ik})^m\ X_k}{\sum_{k=1}^{N} (\mu_{ik})^m} \qquad (7)$$

The norm of distance between i-th cluster and k-th data is :

$$D_{ik}(X_k, V_i) = \frac{\sqrt{\det(F_{mi})}}{P_i} * \exp\left(\frac{1}{2}(X_k - V_i)^T F_{mi}^{-1}(X_k - V_i)\right) \qquad (8)$$

$$F_{mi} = \frac{\sum_{k=1}^{N} (\mu_{ik})^m (X_k - V_i)(X_k - V_i)^T}{\sum_{k=1}^{N} (\mu_{ik})^m} \qquad (9)$$

where the F_{mi} is the fuzzy covariance matrix of the i-th cluster, μ_{ik} is the fuzzy partitioning matrix, m is the weighting exponent controls the 'fuzziness' of the resulting cluster and P_i is aprior probability of selecting the i-th cluster. The distance in Eq. (8) is used in the calculation of P_i, the probability of selecting the i-th cluster given the k-th data point, is given by:

$$P_i = \frac{1}{N} \sum_{k=1}^{N} \mu_{ik} \qquad (10)$$

8 - FUZZY BACKPROPAGATION NETWORK

Artificial neural networks are massively parallel adaptive networks of simple non liner computing elements called neurons which are intended to abstract and model some of the functionality of the human nervous system in an attempt to partially capture some of its computational strengths. Neural networks are classified as feed forward and feedback networks. Back propagation network is of feed forward type. In BPNN the errors are back propagated to the input level[22]. by combine Gath-Geve fuzzy clustering algorithm with back propagation network to produce Fuzzy back propagation neural network, then the fuzzy back propagation neural network (FBPNN) algorithm is as follows[13] [22] [24]:

step 1: create initial random weights for network nodes .

Step 2: A vector pair $[X_p, T_p]$ of the training set, is selected in random.

calculate output for each node in each layer L in network.

$$net_{pj}^{L+1} = \sum_{i=1}^{n^L} w_{ij}^{L} \, out_{i}^{L} + bias_{j}^{L+1} \tag{11}$$

$$out_{pj}^{L+1} = f\left(net_{pj}^{L+1}\right) = \frac{1}{1 + e^{-\beta \, net_{pj}^{L+1}}} \tag{12}$$

Step 3 : calculate the error between actual output out_{pj} and target output, and

use the actual output out_{pj}^{o} with target output t_{pj} to calculate δ

$$\delta_{pj}^{o} = \left(t_{pj} - out_{pj}^{o}\right) f'(net_{pj}^{o}) \tag{13}$$

Step 4: calculate δ value for each hidden layer

$$\delta_{pi}^{L+1} = f'(net_{pi}^{L+1}) \left[\sum_{j=1}^{m^{L+2}} \delta_{pj}^{L+2} \, w_{ij}^{L+1} \right] \tag{14}$$

Step 5: update the weights by adding z_i to the standard update weight equation

for back propagation network, then this equation becomes as follows:

$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}^{L} \tag{15}$$

$$\Delta w_{ij}^{L} = z_i \, \eta \, \delta_{pj}^{L+1} \, out_{pi}^{L} \tag{16}$$

Where $z_i = (\mu_{ik})^m$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(D_{ik} / D_{jk}\right)^{2/(m-1)}} \tag{17}$$

Step 5 : return to step 2 , repeated for each pattern of training set.

9- EXPERIMENTAL AND RESULTS

In this research we used Gath-Geve (G-G) fuzzy clustering algorithm, and combined Gath-Geve with back propagation neural network to produce Fuzzy Back Propagation network (FBPN) algorithm were applied using NSL-KDD data set to classify this data set into 5 classes according to the type of attacks (Normal, DoS, Probe, U2R, R2L). were used "kddTrain+.TXT" file that contain (125973) records and "kddTest+.TXT" file that contain on (22544) records[9]. Table (1) shows the NSL-KDD data set used in training phase, and table (2) shows the NSL-KDD data set used in testing phase that contain from normal and attack connection records[10].

Table 1: The number of samples NSL-KDD data set that were used in training

Type of attack	Kdd Train(NSL) Data set
Normal	67343
Dos	45927
Probe	11656
U2R	52
R2L	995
Total	125973

Table 2: The number of samples NSL-KDD data set that were used in testing

Type of attack	Kdd test (NSL) Data set
Normal	9711
Dos	7458
Probe	2421
U2R	67
R2L	2887
Total	22544

Two approaches G-G, FBPN algorithms was applied on NSL-KDDTrain file data set to classify this data into 5 classes. The result of classification rate (CR%) obtained is 100% in training phase for each of G-G fuzzy clustering algorithm and FBPN algorithm to classify data into 5 classes, one class for normal behavior and 4 classes for different types of attacks. Table (3) shows the Iteration Number (IN) and Time(T) which needed these two algorithms G-G, FBPN in training phase to classify the data.

Table 3: (IN) and (T) for (G-G fuzzy, FBPN) algorithms

Type of Clustering algorithms	Iteration number (TN)	Time second (T)
G-G	21	31.8
FBPN	6	128.379

Table(4) shows the testing results after applying G-G fuzzy clustering algorithm with the detection rate for each attack and normal.

Table 4: The Testing Results Using G-G Algorithm for 5 Classes of NSL Dataset

Type	Input	Output	DR%
Normal	9711	9687	99.753
DoS	7458	7489	99.586
Probe	2421	0.0	0.0
U2R	67	0.0	0.0
R2L	2887	5368	53.782

Detection rate is enhanced when using FBPN for testing the same data set. And this algorithm detect normal, DoS, Probe, and R2L, but it dose not detect U2R attack. Table (5) shows the results of the testing phase of FBPN algorithm.

Table 5: Results of Testing phase of FBPN Algorithm on NSL Dataset for 5 Classes

Type	Input	Output	DR%
Normal	9771	9711	100
DoS	7458	7451	99.906
Probe	2421	2421	100
U2R	67	0.0	0.0
R2L	2887	2961	97.501

And table (6) shows the comparisons between two algorithms G-G fuzzy , and FBPN algorithms for 5 classes. G-G fuzzy clustering algorithm needed (3.1 second) time while FBPN needed (2.9 second) time and (1 iteration) for each of them in testing phase to intrusion detection.

Table 6: Comparison between G-G fuzzy clustering and FBPN Algorithms for 5 Classes of NSL Dataset

Performance measure	G-G	FBPN
Normal detection	9687	9711
Attack detection	10345	12759
Detection rate_normal	99.753	100
Detection rate_attack	76.684	99.136
Detection_rate	88.219	99.568
Times	3.1 second	2.9 second
Iterations	1	1

CONCLUSION

In this research Gath-Geve (G-G) fuzzy clustering algorithm and neural networks (Fuzzy back propagation network) were applied to classify NSL-KDD data set into 5 classes one for normal behavior and others for types of attacks, and these two algorithms satisfied very good results in classification and detection. The applied approaches (G-G FBPN) algorithms improved a high classification rate 100% in training phase. And the application of these approaches made the intrusion analysis engine more simple and efficient. These two algorithms obtained a high detection rate for NSL-KDD dataset. It has been found that FBPN algorithm is the best from G-G fuzzy clustering algorithm.

REFERENCES

[1] Urvashi M., Anurag J., "AN IMPROVED METHOD TO DETECT INTRUSION USING MACHINE LEARNING ALGORITHMS", Informatics Engineering, an International Journal (IEIJ), Vol.4, No.2, June 2016.

[2] Panda M., Patra M., "some clustering algorithms to enhance the performance of the network intrusion detection system " ,journal of theoretical and applied information technology, pp.795-801, 2008.

[3] Gomez J., Dasgupta D., “Evolving Fuzzy Classifiers for Intrusion Detection”, proceeding of the 2002 IEEE.

[4] Song D., Heywood M., Zincir-Heywood A., “Training Genetic Programming on Half a Million Patterns: An Example from Anomaly Detection”, IEEE Transaction on evolutionary computation, 2005.

[5] Liu X., Tian J., "Application research of improved K-means algorithm in intrusion Detection" International Conference on Computational Science and Engineering pp.96-100 (ICCSE 2015).

[6] Vemuri V., “ENHANCING COMPUTER SECURITY WITH SMART TECHNOLOGY”, TK5105.59.E62, 2005.

[7] Sabnani S. V., “Computer Security: A Machine Learning Approach”, Royal Holloway, University of London, 2008.

[8] Chimphlee W., Abdullah A., Sap M., Chimphlee S., Srinoy S., 2007, "A rough-fuzzy Hybrid Algorith for Computer Intrusion Detection ", the international Arab journal of information Technology, Vol.4, No.3.

[9] Chebrolu S., Abraham A., Thomas J., "Feature deduction and ensemble design of intrusion detection system ", www.elsevier.com/locate/cose . 2004.

[10] Siddiqui M., "high performance data mining techniques for intrusion detection ", thesis 2004.

[11] Al-Sharafat W., Naoum R., “Adaptive Framework for Network Intrusion Detection by Using Genetic-Based Machine Learning Algorithm”, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, 2009.

[12] Lakhina S., Joseph S., verma B., “Feature Reduction using Principal Component Analysis for Effective Anomaly-Based Intrusion Detection on NSL-KDD”, International Journal of Engineering Science and Technology Vol. 2(6), 2010.

[13] http://nsl.cs.unb.ca/NSL-KDD/
Ramamurthy N., Varadarajan S., "Roubust digital image watermarking scheme with neural network and fuzzy logic approach", international journal of emerging technology and advanced engineering, Vol. 2, Issue 9, pp 555-562, 2012.

[14] Faraoun K. M., Boukelif A., “Neural Network Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusion”, International Journal of Computation Intelligence 3;2 2007.

[15] Betanzos A., Marono N., Fortes F., Romero J., Sanchez B., “Classification of computer intrusions using functional networks. A comparative study”, ESANN, European Symposium on Artificial Neural Networks, 2007.

[16] Zhang C., Jiang J., Kamel M., 2005, “Intrusion detection using hierarchical neural networks”, Pattern Recognition Letters 26, 779-791.

[17] ENGEN V., 2010, “Machine Learning For Network Based Intrusion Detection”, Ph.D. thesis, Bournemouth University.

[18] Yan K., Wang S., Liu C., 2009, “A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks”, International Multi Conference of Engineers and Computer Scientists, Vol. I

[19] Vlad Z., Ofelia M., Maria T., " fuzzy clustering in an intelligent agent for diagnosis establishment", Scientific Bulletin of the Petru Maior University of Tirgu Mures, Vol. 6, ISSN: 1841-9267, 2009.

[20] Hasanzadeh R., Moradi H., Sadeghi H., " fuzzy clustering to the detection of defects from nondestructive testing", International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, March 27-31, TUNISIA, 2005.

[21] Yi Xing Z., Zhang Y., Hou Y., Jia L., " on generating fuzzy systems based on pareto multi-objective cooperative coevolutionary algorithm", International Journal of Control, Automation, and systems, Vol. 5, No. 4, pp 4444-445, 2007.

[22] Sarkar D., "Methods to speed up Error Back-propagation learning Algorithm ", ACM Computing Surveys, Vol. 27, No. 4, pp 519-541, 1995.

[23] Wasserman P. D., "Neural Computing theory and practice", New York, 1989.