International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Detecting Intrusion with High Accuracy: using Hybrid K-Multi Layer Perceptron

Amit Dogra, Taqdir

Abstract: ***The intrusion detection is the mechanism by which abnormality from the state driven dataset is discovered. The intrusion causes the problem of false discovery that mislead overall result. The resources from server may not be accessed by the use of intrusion be malicious users. The propose mechanism of Self organizing KMLP technique to discover abnormal patterns from the dataset. The dataset is synthetically derived to demonstrate the experimental work. The operation is demonstrated against K-Map clustering. The result is presented in terms of classification accuracy, number of attacks and execution time and result shows significant improvement by the margin of 10%.***

Keywords : intrusion detection, machine learning , KMLP

I. INTRODUCTION

The detection of malicious action from within network is known as intrusion Detection system. The attack to confidentially, availability and integrity of network and the user is known as intrusions. Today there is need to detect intrusion as there is large set of data exist and this data suffers from missing values. The data values can be missing due to incorrect measurements, manual data entry procedure and finding missing data is difficult task. To handle data efficiently and detecting intrusion into the network the use of machine learning algorithm along with pre-processing mechanism is used.

The intrusion within the dataset causes significant effect on the classification accuracy. [1]The pattern discovery is hampered by the presence of intrusion within the dataset. The detection of abnormality within dataset is researched over by the use of clustering mechanism. [2]The clustering mechanisms that are commonly employed along with advantages and disadvantages are discussed in this literature. In addition attacks that are common on dataset are also elaborated through this literature. [3]The attacks on dataset could be on resources, network or data.

Manuscript published on 30 September 2019
* Correspondence Author
Amit Dogra*, Assistant Professor, Department of Computer Science and Engineering at SoET, BGSB University Rajouri (J&K).
Dr. Taqdir, Assistant Professor, Department of Computer Science and Engineering, Guru Nanak Dev University, R/C Gurdaspur.

The attacks on resources causes denial of service and hence extra time is consumed while accessing resources. These attacks if severe causes starvation problem. The fuzzy attacks are node based upon the membership functions and if membership functions if abnormal can cause attacks in the dataset. Multiple identity attack is also a problem in which receiver could not determine the sender or vice versa. In all the situations attack hamper the performance of the system considered. This literature focuses on detection of attacks and high degree of classification accuracy.

The paper is organized as under: section 2 gives the comparative analysis of intrusion detection mechanisms to explore parameters that can be optimized in this research, section 3 gives the indepth analysis of proposed system, section 4 gives result and performance analysis and section 5 gives conclusion and future scope.

II. LITERATURE SURVEY

This section gives the in-depth into the work that is done towards discovering intrusions using distinct mechanisms of machine learning. To this end optimization along with machine learning algorithms are widely considered for evaluation.

In [4] proposed Fuzzy C mean clustering algorithm for detection of attack within the network. The prediction accuracy of the system is high. It describes clustering algorithm that are based on the fuzzy rule set. The execution time is considerably high in this system.

[5]describes technique that is based on Artificial bee colony algorithm for cluster formation. It uses cluster head selection for detecting the malicious node entry into the network. The optimization of selection is done but the false negative rate is very high.

[6] proposes Heuristic clustering based approach that handles intrusion on the basis of entry of node. The accuracy of the system is high as it efficiently selects the class of node. But the error rate is high as there many malicious node in the clusters that are complex to detect.

[7] describes a technique that detect the intrusion within the network using coverage rate. This technique is called coverage aware clustering. It first of all detects the nodes that are likely to be attacked using cluster head selection technique. The prediction of cluster head selection is high but the error rate is more.

[8] proposes Classification and clustering algorithm that minimized the error rate for selection of cluster head. The error rate is handled efficiently and intrusion is detected faster. But the relation degree between nodes and segments is maximized.

[9] proposes Intrusion detection using feature selection and k means clustering algorithm. This is based on analyzing nearest neighbor and the finding the node that are attacked. The accuracy of predicting attacked node is improved .
[10] proposes intrusion detection system that is based on KDD. It analysis the traffic patterns to detect the packets that could have virus or harm the network. It also handles the missing values present into dataset. The classification accuracy is significantly high.
[11] describe a model for intrusion detection that is based on machine learning algorithms named Random forest and SVM. It reduces the dimensionality of the dataset that makes feature selection process faster. The accuracy is high but complexity is more.
[12] proposes IDS that prevents IOT systems from attacks. The modules of this system analysis sensor hub and recognize the denial of services attack. The proposed systems accuracy is very high and it detects the attack within the IOT system efficiently.
[13] gives a comparative analysis of various machine learning strategies for network intrusion detection. This analysis describe about feature detection and classification algorithms that are used for intrusion detection.

The comparative analysis of techniques used in literature section is presented in table 1

Approach	Technique used	Advantages	Disadvantage
Genetic Algorithm	It describe the using GA for IDS the proposed System is GA rulebase system which including crossover, mutation, fitness, selection process & finally generate the rules for the test	• Has inherent parallelism in its search method. • Able to solve the Optimization problems that can be described with chromosomes encoding	• No guarantee to find best solution ,so sometimes has difficulties to find exact global optimum
Artificial Immune System	It is able to distinguish between good cells and potentially harmful ones .Artificial Immune System are learning and problem solvers based on our immune systems	• Immediate Protection	• No long term protection
Fuzzy Logic [1]	Fuzzy Logic is designed to handle vague and imprecise data. To indicate an intrusion, a relationship between input and output variables is defined by creating different set of rules. It uses membership functions to examine the intensity of truthfulness	• Able to pull the outcome and procedure response in terms of vague, imprecise & defective quality • Use linguistic variables.	• Member Function estimation is hard. • Difficult to create a model for fuzzy logic
Artificial Neural Network	It is used to detect computer attacks, it have typical characteristics of users .In this incoming data is filtered by neural network for suspicious events.	• It is inherently parallel, so is able to operate very effectively. On parallel hardware. • When one of its neurons fails, it still keeps on with no problem.	• It is very hard to say that how (well or unfavorable) is the quality of trained neural network. • Not able to be retrained.

Hybrid approach	reduces the dimensionality of the dataset that makes feature selection process faster	• Gains flexibility & increased security. • In this we find out the unknown attacks and to raise the detection rate.	• Lower false positive and false negative. • Relatively high cost.

Table 1: Comparative analysis of intrusion detection mechanisms.

Next section provides detail of technique used in proposed system to achieve optimality in terms of feature selection and classification.

III. PROPOSED SYSTEM

The proposed methodology consist of three phases : Normalization, feature selection and Classification
These phases are critical for improving classification accuracy of the attack detection. Overview corresponding to these phases is elaborated as under

- Normalization: In this the data is scaled to have a values between 0 and 1. In this process the data set values have been reduced according to the mean and then it is adjusted according to the standard deviation.
- Feature Selection: The Filter Method utilizes measurable ways to deal with give a positioning to the features. After giving the positioning client can pick best k feature for the test like top 24or 41 as indicated by their prerequisite. The process of feature selection is done using multi-layer perceptron algorithm. This algorithm consist of three layers : input layer, hidden layer and output layer. In the commutative frequency is calculated that decides which features are going to be selected.
- Classification: After that hybrid mechanism of adaptive boost , random forest and KNN algorithm is used for classification .In this the features are classified and then evaluation is done.

The algorithms and mechanism implemented to achieve attack detection from within dataset is given in section 3.1, 3.2 and 3.3.

1.1 Normalization

The Normalization mechanism is followed in order to ensure data can be handled successfully. This phase is necessary since KMLP model cannot accommodate numeric data of varying size. This distance based approach can accommodate precision upto 5 digits only. The algorithm used to perform normalization in KMLP is given as under
Algorithm(Normalization)

- B=N // N is the data fetched from dataset
- C=floor(B)// C gives the floor value corresponding to decimal digit
- For i=0: B>=1;B++
  B=B/10 // Used to count number of digits
  End of for
- D=N-i // i contains length of data and N is the actual distance value from dataset
- B=N // Re-initialization
- B=B*power(10,D)
- E=b+0.5 // Rounding off number in case value exceeds 0.5
- If E=ceil(B)
  F=ceil(B)
  H=F-2;
  If(H%2!=0)
  E=E-1
  End of if
  End of if
- J=floor(E)
- M=power(10,D)
- J=J/M
- Output J

'J' Contains actual round off value for the distance value that is used during training operation. Training operation through distance value allows fixed value to be obtained that can be used to determine attacks.

1.2 Feature selection

Feature selection is the primary phase that is used to determine the abnormal patterns from the dataset. The abnormal pattern detection is distance based. The obtained values from the dataset is clustered using KNN mechanism. The population in terms of clusters is fed into multi-layer perceptron. The MLP includes three layers including input , processing and output layer.

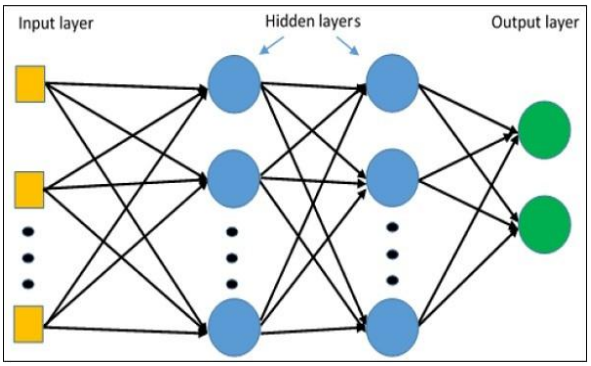

Figure 1: Multi-layer perceptron mechanism

KNN approach is used to determine nearest neighbor to form cluster. The Euclidean distance is used to evaluate distance and value of K is fixed to 5.

$$ED = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

Equation 1: Euclidean distance for clustering.

The clustered data is fed into MPL model. Two classes indicating maliciousness and non maliciousness. Hidden layer is also known as processing layer containing weight factor. The weight factor is adjusted until the data lie within the specific class. The process of MLP is iterative in nature. The iterative approach plays critical role in achieving optimality. The class so predicted is fed into output layer. The algorithm for the KMLP approach is given as under

- Input layer definition

$$Nodes_i = Dataset_{Traffic_{tupples_i}}$$

where i define number of tuples within the traffic da

- Hidden layer definition

Processing layer contains the neurons which are defined by normalization function

$$Normalization_i = \frac{Total_{Nodes}}{K}$$

Where, K is the parameter while value vary within the range of 2 to 4.

- Activation function definition

Activation function indicates the activation of weight function which is gradient descent in this case.

$$update_i = lerning_{rate} * gradient\_of_{parametes}$$

learning rate is constant for each iteration

- Backpropagation for weight adjustment

This is used to check prescribed tolerance. If it is not achieved then weight is adjusted by the random factor between [1-5].

If (Perscribed_tolerance)

Then list error rate

Else

Adjust weights $w_{i_j} = W_{i_j} + rand(1,5)$

1.3 Classification

This phase is used to determine the class in which obtained value from MLP lies. The classification mechanism involves random forest, adaptive boost and KNN approach. The classification phase using the hybridization of random forest, adaptive boost and KNN is given as under

- Initialization

Clusters data from the KNN is served as initial population.

- Fitness function

Fitness function involves minimization of errors obtained after MLP

If (Minimum_Error)

Convergence with prediction and error rate

Else

Move to the next phase

End of if

- Based on fitness function, selection of population
- Performing crossover

2 or 4 point crossover is performed and nomination is obtained

$$Normalization_i = \frac{Normalization_i}{K}$$

Normalization selects the parents for next phase

- Performing Change over

Changeover the initial data and new nodes are generated.

The classification phase gives the result of the proposed mechanism in terms of execution time and reliability. The detailed methodology is given in figure2.

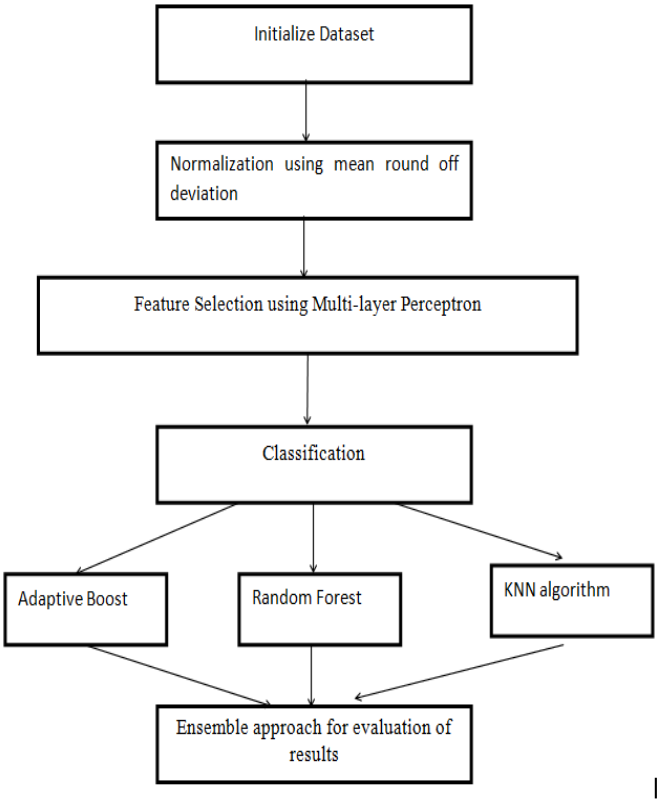

Figure 2:Methodology of proposed system

IV. PERFORMANCE ANALYSIS AND RESULT

The result of the proposed system is given in terms of classification accuracy and execution time. The classification accuracy is obtained using rate of positive detections to the total malicious detections. The classification accuracy result is predicted from synthetic dataset. The obtained result with the application KMLP and hybrid classification yield 10% betterment in terms of classification accuracy. The equation 2 gives the classification accuracy used in proposed system.

$$CA = \frac{True_{Predicted}}{Total_{Prediction}}$$

Equation 2: 'CA' classification accuracy

The result of the proposed approach is compared against plain MLP approach and Random forest approach. MLP approach yield better result in terms of prediction but execution time is sufficiently higher as compared to proposed approach.

The result of classification accuracy is given in figure 3. Simulation is tested against 100,200 and 300 data points. In every case result obtained is better.

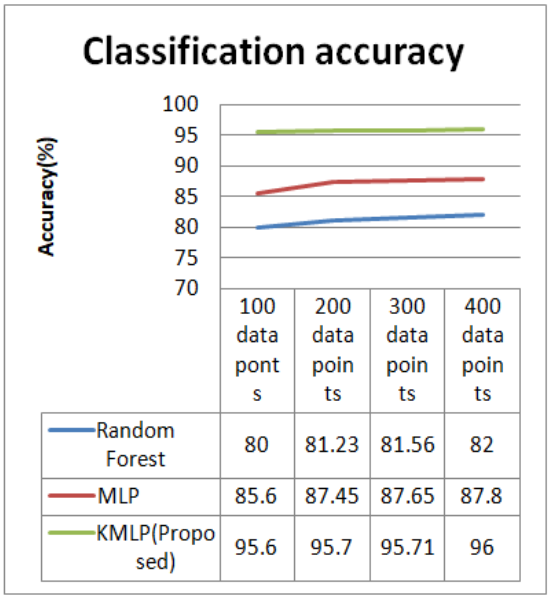

Figure 3: Classification accuracy

The data point increments does not alter classification accuracy much. Although classification accuracy is highest at 400 data points.

Execution time is another key parameter that is going to decide the success of proposed system. The execution time is sufficiently reduced by the application of normalization and KMLP. The hybridization of classification algorithms such as random forest, adaptive boost and KNN gives better result and increases reliability. Execution time increases as the number of data points increases but it is less than random forest and sole MLP mechanism as shown in figure 4

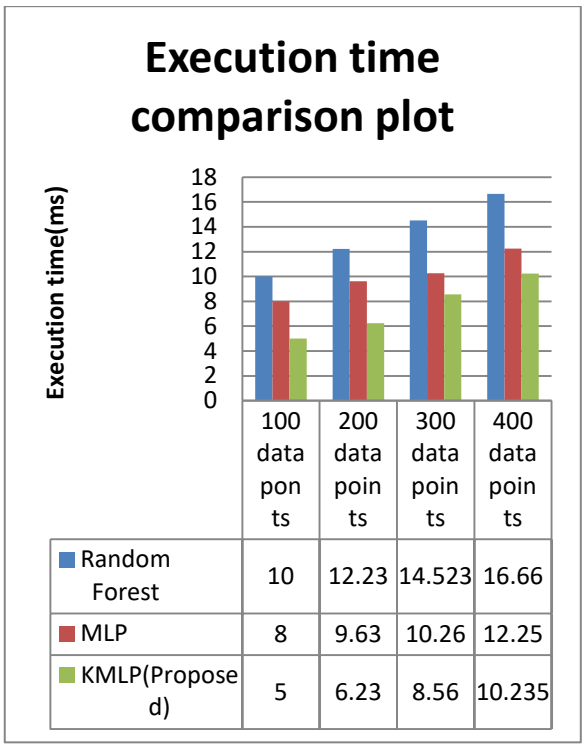

Figure 4: Execution time comparison plot

The result obtained is better in case of Normalized KMLP approach with hybrid classification model. The result is demonstrated against the synthetic dataset. Real time dataset can be obtained from UCI or kaggle website and proposed model testing could be done in future.

V. CONCLUSION AND FUTURE SCOPE

The approach followed in proposed system first of all normalize the data by following round off procedure to fit the data within framework. After normalization, KMLP approach is applied to distinctly form clusters and then feeding operation take place. This feeding operation fed the clusters into MLP model. The MLP model bring the data into compatible zone classes. The classification process is based on hybridization of random forest, adaptive boost and KNN approach. The classification accuracy is higher in case of proposed system. Execution time is also decreased in the detection process. The execution time and classification accuracy is considered primary parameters to determine validity of proposed approach. The experiment is conducted on synthetic data and result is better.

In future real time dataset can be tested against the proposed system and validity can be examined. In addition confusion matrix can be accommodated to examine more parametric results.

REFERENCES

1. S. Muhammad, S. Hussain, and M. Yousaf, “Neighbor Node Trust Based Intrusion Detection System for WSN,” *Procedia - Procedia Comput. Sci.*, vol. 63, pp. 183–188, 2015.
2. Q. Yan, F. R. Yu, Q. Gong, and J. Li, “Software-defined networking (SDN) and distributed denial of service (DDOS) attacks in cloud computing environments: A survey, some research issues, and challenges,” *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 602–622, 2016.
3. N. Alsaedi, F. Hashim, and A. Sali, “Energy Trust System for Detecting Sybil Attack in Clustered Wireless Sensor Networks,” *IEEE Access*, no. Micc, pp. 91–95, 2015.
4. Z. Sun, L. Gao, S. Wei, and S. Zheng, “A Fuzzy C-Means Clustering Algorithm and Application in Meteorological Data,” *s2010 Second Int. Conf. Model. Simul. Vis. Methods*, pp. 15–18, 2010.
5. D. C. TRAN Dang Cong, WU Zhijian, WANG Zelin, “A Novel Hybrid Data Clustering Algorithm Based on Artificial Bee Colony, Algorithm and K-Means,” *Chinese J. Electron.*, vol. 24, no. 4, 2015.
6. J. Zhao, K. Yang, X. Wei, Y. Ding, L. Hu, and G. Xu, “A Heuristic Clustering-Based Task Deployment Approach for Load Balancing Using Bayes Theorem in Cloud Environment,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 305–316, Feb. 2016.
7. F. Awad, E. Taqieddin, and A. Seyam, “Energy-Efficient and Coverage-Aware Clustering in Wireless Sensor Networks,” *acm*, vol. 2012, no. July, pp. 142–151, 2012.
8. S. Kiruthiga, “Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques,” *IEEE*, 2014.
9. A. Sharma, “Spam Filtering using K mean Clustering with Local Feature Selection Classifier,” *ijca*, vol. 108, no. 10, pp. 35–39, 2014.
10. D. Shona and M. Senthilkumar, “An ensemble data preprocessing approach for intrusion detection system using variant firefly and Bk-NN techniques,” *Int. J. Appl. Eng. Res.*, vol. 11, no. 6, pp. 4161–4166, 2016.
11. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, “An Investigation on Intrusion Detection System Using Machine Learning,” *Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018*, no. January, pp. 1684–1691, 2019.
12. V. Pandu, J. Mohan, and T. S. P. Kumar, “Network Intrusion Detection and Prevention Systems for Attacks in IoT Systems,” no. January, pp. 128–141, 2019.
13. Y. Hamid *et al.*, “Machine Learning Techniques for Intrusion Detection : A Comparative Analysis To cite this version : HAL Id : hal-01392098 Machine Learning Techniques for Intrusion Detection : A Comparative Analysis,” 2016.

AUTHORS PROFILE

Dr.Taqdir
Designation: Assistant Professor.
Qualification: PhD (CSE, NIT Jalandhar),M.Tech(CSE,),B.Tech(CSE).
Department: Computer Science and Engineering.
Email: taqdir_8@rediffmail.com.

Dr. Taqdir is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Guru Nanak Dev University, R/C Gurdaspur. She holds the experience of 17 years in teaching at University level and her area of interest is Digital Image Processing and Machine Learning.

Mr. Amit Dogra
Designation: Assistant Professor.
Qualification: M.Tech, B.E.
Department: Computer Science and Engineering.
Email: amitdogra004@gmail.com.

Amit Dogra is currently working as an Assistant Professor in the Department of Computer Science and Engineering at SoET, BGSB University Rajouri (J&K). He holds a B.E degree in Information Technology and M.Tech degree in Networking and Internet Engineerin