Framework for Providing Security in Private Cloud using Machine Learning Techniques

Shridhar Allagi, Rashmi Rachh

Abstract: ***The advancement in cyber-attack technologies have ushered in various new attacks which are difficult to detect using traditional intrusion detection systems (IDS).Existing IDS are trained to detect known patterns because of which newer attacks bypass the current IDS and go undetected. In this paper, a two level framework is proposed which can be used to detect unknown new attacks using machine learning techniques. In the first level the known types of classes for attacks are determined using supervised machine learning algorithms such as Support Vector Machine (SVM) and Neural networks (NN). The second level uses unsupervised machine learning algorithms such as K-means. The experimentation is carried out with four models with NSL- KDD dataset in Openstack cloud environment. The Model with Support Vector Machine for supervised machine learning, Gradual Feature Reduction (GFR) for feature selection and K-means for unsupervised algorithm provided the optimum efficiency of 94.56 %.***

Keywords: Intrusion Detection System (IDS), Support Vector Machine (SVM), Supervised Machine Learning, Unsupervised Machine Learning, Gradual Feature Reduction (GFR).

I. INTRODUCTION

Recent advancements in the field of cloud computing, big data and Internet of things (IoT) have contributed a lot of digital data floating around the communication world. The organization's business credibility depends on the security of the data across its various nodes and preserving the privacy of the shared data among set of authorized users [1]. The various traditional methods implemented by organizations to secure the network by writing the static rule sets for access control using various firewalls. The recent advancement in the field of cyber hacks have enforced the organizations to shift the security policies from firewalls systems to the advanced dynamic systems.

Evolution in the field of machine learning has attracted various organizations and researchers to implant the advanced machine learning algorithms in the field of network security. The capability of machine learning models to analyze various types of cyber-attacks by using the network log data has really enhanced the confidence in security systems.

This paper is organized as follows. Section I gives brief introduction. In section II reviews of the related work is summarized. Section III provides details of the proposed model. In section IV, the results of the proposed model are discussed and Section V concludes the paper.

1.1 Network Security

Network security in reference to organizations refers to preserving the integrity, confidentiality and availability of data across the clock. Any attempt to tamper the above three aspects is attempted to be a security breach of the system. The attacks to the organization network can be various types [2]. The three general categories of attacks are 1) the intruder gets unauthorized access to network, and steal the information related to the organization without changing the state of the system. 2) The intruder manipulate the performance of the service by degrading the system performance (DoS). 3) The intruder gets the unauthorized access to the system and adulterate the information in the organization system.

The prominent security systems for the organizations are Intrusion detection Systems (IDS). These systems are located in premises of the organization and regularly examines the incoming and outgoing traffic of the networks and any anomaly reported will be alerted to the administrator system for further actions. The IDS located in networks are Network Intrusion detection systems (NIDS) and the IDS for hosts are Host Intrusion Detection Systems (HIDS). In our work, we focus on NIDS.

1.2 Machine Learning

Machine learning being the advanced field of study in engineering accelerates the capacity of self-learning of the machine without being explicit programming. The machine learning can be defined as the learning capacity L of Machine M builds with the set of tasks T with the time t. The rate of L increases with time t and set of more tasks T. The machine learning is broadly classified into supervised and unsupervised machine learning.

1.3 Supervised Machine Learning

Supervised machine learning is a field of ML, where predicating variable X' are labelled to particular class Y' by model which is trained with labelled dataset. The selection of the predicting variable plays a crucial role for the efficiency of model. The widely popular supervised machine learning algorithms are support vector machines and neural networks which are basically implanted in regression and classification problems.

The support vector machine sorts out the input data into two categories, on the principal of separating the hyper planes over the higher dimension data reducing to the hyper planes of few labels or classes. SVM's serves the best optimal solutions for performing the liner classification using the probabilistic hyper plane structures. The fig. 1 shows the sample SVM classifier for two optimized hyper planes reducing the higher dimension data.

Revised Manuscript Received on October 10, 2019.

***** Correspondence Author

Shridhar Allagi, Department of Computer Science and Engineering, KLE Institute of Technology, Hubbali, India, Email: shridharallagi1@gmail.com

Dr. Rashmi Rachh, Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, India

Email: rashmirachh@gmail.com

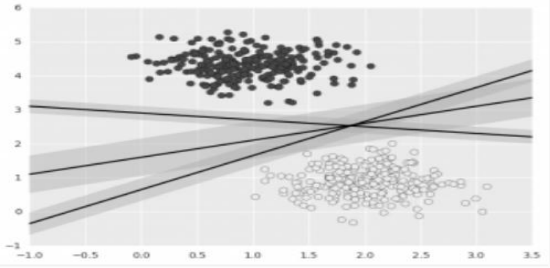


Fig.1 : Support Vector Machine for two Hyper planes [2]

The hyper planes are generated using the function kernal. The different significant kernal functions are sigmoid, radial , linear and polynomial. The circular seperating hyper planes are signified by radial and sigmoid.

The neural networks is another trending supervised machine learning where it works on the concept of brain mapping with neurons representing the nodes in the NN graphs. The most significant models used in NN are gradient descent and conjugate gradient.

1.4 Unsupervised Machine Learning

The unsupervised Machine learning is a self-organizing model where the predicating variables X' are not entered with any preexisting labels. The model iteratively learns about the undiscovered patterns in subsequent iterations. The most prominent unsupervised machine learning for clustering is k means algorithm.

The K means works on the principal of distance measurements and reduces the n high dimensionality data in to pre-configured k centroids.

1.5 Feature Selection in Machine Learning

The feature selection process plays a vital role in implementing the machine learning models. The feature selection methods works for dimensionality reduction where only the dimensions of importance are preserved. Our model works on the popular feature selection techniques Gradual feature reduction (GFR) and principal component analysis (PCA).

The gradual feature reduction works on the principal of not neglecting any of the features of their importance. It takes in account of all the features. It works by eliminating the one feature and calculates the performance of model. Similarly in the next iteration it eliminates other feature and calculates the model performance. It iteratively performs for all the set of features and finally considers the set of features with highest efficient model.

The principal component analysis (PCA) is implemented on the factor of reducing n dimension data into p dimensions, ensuring that entire dataset is described with p new attributes. The PCA indicates the low dimension representation of dataset. The covariance matrix is created and the p dimensions are sorted in decreasing order of covariance values.

1.6 Performance Metrics

The popular performance factors in machine learning are to create the confusion matrix. The derived basic matrices such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are determined and can be graphically visualized using the confusion matrix.

Using the matrix following performance measures are computed.

- Accuracy: Percentage of entries classified correctly by the model.

  Accuracy = (TP+TN) / N (1)
- Error : Describes the incorrectly classified entries

  Error = (100 – Hit Rate) (2)
- Precision= Percentage hits over entries of positive class that were classified as positive class.

  Precision = (TP)/ (TP+FP) *100 (3)
- F- measure: Derived effectiveness measurement.

  F-measure= 2* (Precision * recall)/ (Precision + recall) (4)

II. RELATED WORKS

In this section, review of the work carried out in the areas of Machine learning and network security is discussed.

W.feng et. al [3]have proposed a hybrid classification machine learning algorithm. The model classified the network log data into normal or abnormal using support vector machine in earlier stage and clustering at later stage using self-organized any colony method. The method was useful for pre-defined patterns of attack but lacks to detect the new unknown patterns resembling the attacks.

D. M. Farid and M. Z. Rahman [4] have focused on classification of alerts to reduce the number of false postives in IDS using self-adaptive Bayesian algorithms. The model predicts the class to which attack belongs using the limited computational resources. The model achieves better efficiency for less number of classes and performance decreases with computational complexity with increase in number of classes.

J. Jha and L. Ragha [5] have proposed a model for classification of attacks using the hybrid feature selection method which combines the wrapper and filter methods. All the probabilistic feature sets were ranked with information gain. The feature set with highest efficiency was considered for classification of types of attacks. The computational complexity was overhead to determine the information gains for all feature sets. The new types of attacks with additional features were biased towards normal attacks.

J. F. Joseph and Das [6] have focused on sinking behavior in ad hoc networks using autonomous host based IDS. The detection accuracy was maximized using the cross layer approach. The accuracy was further enhanced using the support vector machines. However computational is expensive for SVM. To reduce the cost, model uses the linear classification algorithm Fischer discriminants analysis (FDA) to eliminate the data with low entropy.

KyawThetKha [7] have proposed the enhanced SVM model with Recursive feature elimination (RFE) and k-means to assign ranking to features and selection of features with highest ranks. The model used the supervised machine learning algorithms and works on pre-defined attack patterns.

From the comprehensive literature review, it can be concluded more of work was focused on supervised machine learning algorithms which detected the trained patterns, but failed to capture new patterns resembling attacks. The proposed work focuses on using both supervised and unsupervised machine learning algorithms to detect pre identified patterns of attack and also new patterns resembling attacks.

III. PROPOSED MODEL

The proposed model uses the combinations of supervised machine learning algorithms Support vector machine and neural networks with unsupervised machine learning k-means. The model is implemented using the principal component analysis (PCA) and Gradual feature reduction (GFR) for feature selection. The NSL –KDD dataset is used to perform the experimentation.

In this paper, two level security framework is proposed which uses machine learning techniques to detect newer attacks. In our work, we propose the two level dynamic machine learning algorithms and define the model with the highest efficiency to detect the anomalies in the network intrusion detection systems (NIDS).

Table I gives the different combinations of machine learning algorithms implemented in our work.

TABLE I: Combinations of Algorithms in Proposed work

Models	Supervised ML	Feature Selection	Unsupervised
Model 1	SVM (Support Vector Machine)	PCA (Principal Component Analysis)	K-means
Model 2	NN (Neural Networks)	PCA (Principal Component Analysis)	K- means
Model 3	SVM (Support Vector Machine)	GFR (Gradual Feature Reduction)	K-means
Model 4	NN (Neural Networks)	GFR (Gradual Feature Reduction)	K- means

The configuration for NN used was one hidden layer with parameters {i, n, j} representing the number of neurons in input, hidden and output layers.

The SVM cost/gamma function with default optimal values were used for the model. The parameter value for K-means was fixed to 2 clusters.

The revised data set of KDD99 with eliminated duplicate records i.e. NSL_KDD dataset was used in the experimentation. The four configuration models efficiency was computed with respect to the efficiency parameters defined in the earlier section.

The fig.2 shows the steps involved in our methodology

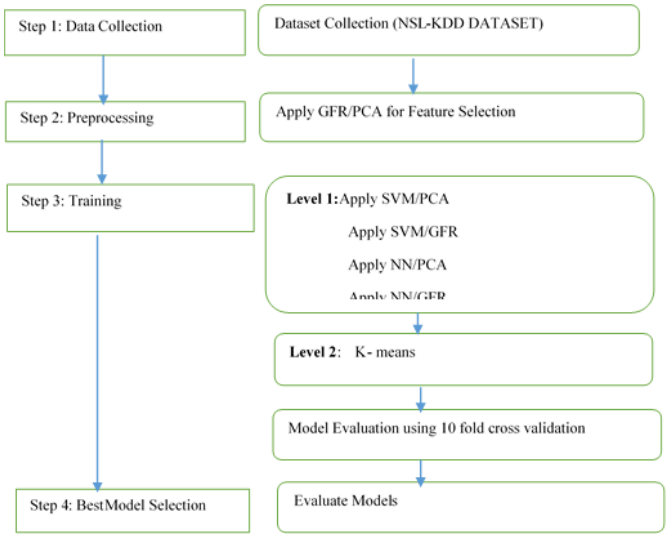


Fig. 2 Workflow of Model for training

Assuming all the attributes contribute equally to the dataset, we use the k fold cross validation method for identifying the best suitable model. The configuration is set to k=10 (i.e each data is split to 10 folds).

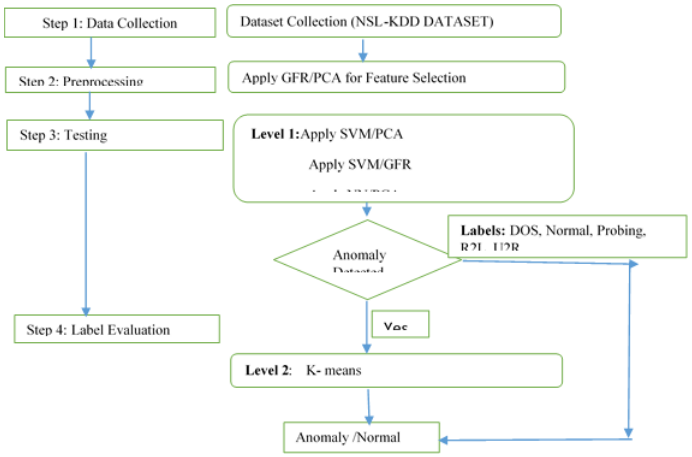


Fig.3 Workflow for testing

Fig.3 shows the model for the testing the dataset. The dataset is first subjected to the preprocessing step, where the various features are extracted with the GFR and PCA method. After feature extraction, first level of supervised learning algorithms are applied and the one with the label are passed to the second level of unsupervised k- means clustering, where it clusters the data to normal or abnormal The two level of supervised and unsupervised machine learning algorithm enhances the performance by reducing the computing complexity by subjecting the data detected as anomaly in first phase to unsupervised.The supervised machine learning showed the significant performance in the network intrusion detection systems (NIDS) for the known type of attacks and vulnerabilities. In order to detect new type of attacks without pre-defined patterns, the efficacy of supervised machine algorithms falls short and unsupervised algorithms are found appropriate.

IV. RESULTS AND DISCUSSIONS:

Dataset: The NSL KDD dataset was used for experimentation. The dataset consisted of four labelled classes DoS, Normal, R2L, U2R and Probe. The NSL KDD dataset is revised version of KDD, where the enormous redundant data is eliminated. The dataset had captured the various interactions with the network with various 41 features. The entire dataset comprises of 39 types of different attacks categorized to four classes DoS, U2R, R2L and Probe. The table 2 shows list of various attacks grouped to classes.

TABLE 2: Various attacks classified to four classes

CLASS NAME	ATTACKS
R2L	ftp_write , imap, phf, warzmaster, multihop, xlock, snmpguess, xsnoop, snmpgetattack, sendmail, named, httptunnel, warzclient
U2R	Loadmodule, rootkit, perl, sqlattack, xterm, Ps, buffer_overflow
Probe	Nmap, portsweep, saint, Ipsweep, satan,Mscan
DoS	Neptune, pod, Land, Back, Apache2

The dataset is captured across the TCP, UDP and ICMP protocols. The dataset attributes labels are renamed to Normal: 0 , Dos :1, U2R : 2, R2L :3 , Probe : 4. The values are normalized with

$X' = ((X+\mu) / \sigma)$ where X' is normalized value, X is captured value, μ is mean and σ is standard deviation.

The features computed in PCA and GFR are represented in fig. 3. The experiments were carried out in the Openstack Newton. Python 3 was used with sciket learn and tensorflow.

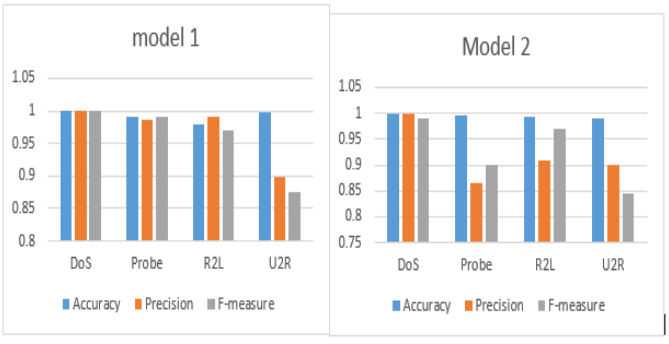
```
Features selected for DoS: ['logged_in', 'count', 'serror_rate', 'srv_serror_rate', 'same_srv_rate',
'dst_host_count', 'dst_host_srv_count', 'dst_host_same_srv_rate', 'dst_host_serror_rate', 'dst_host_s
rv_serror_rate', 'service_http', 'flag_S0', 'flag_SF']

Features selected for Probe: ['logged_in', 'rerror_rate', 'srv_rerror_rate', 'dst_host_srv_count', 'd
st_host_diff_srv_rate', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'dst_host_rerro
r_rate', 'dst_host_srv_rerror_rate', 'Protocol_type_icmp', 'service_eco_i', 'service_private', 'flag_
SF']

Features selected for R2L: ['src_bytes', 'dst_bytes', 'hot', 'num_failed_logins', 'is_guest_login',
'dst_host_srv_count', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'service_ftp', 's
ervice_ftp_data', 'service_http', 'service_imap4', 'flag_RSTO']
```

Fig. 3 Features computed with PCA and GFR

The confusion matrix is generated for each model and the performance factors discussed in earlier sections are computed. Figure 4represents the performance factors computed.



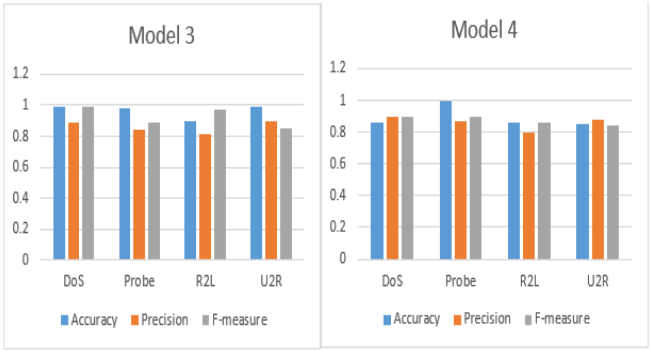


Fig. 4 Performance of Various models

The analysis of various models across the dataset suggests that the model1 (SVM and PCA)are suitable for detection of attacks such as DoS, probe and R2L. The model fails to detect the U2R anomalies. The model 2 (NN and PCA) shows the good result for DoS but average for probe and U2R. The model 3 (SVM and GFR) achieves the highest rate of accuracy in all types of labelled attacks. The model 4(NN and GFR) performs comparatively better than model 1 and 2 but triggers more false alarms compared to model 3.

V. CONCLUSION

In this work, a security framework has been implemented to provide secured intrusion detection system for the private cloud by harnessing the capabilities of supervised and unsupervised machine learning algorithms. The model uses SVM and NN in first phase with PCA/GFR for feature selection technique and uses the k-means in second level of unsupervised machine learning to detect the new attacks in the network. The work demonstrates that GFR feature selection is more suitable for known and unknown attacks in the network. GFR with SVM and K means achieves the highest efficiency up to 94.56 %. The work can be further extended with reinforcement learning to reduce the false alarms.

REFERENCES

1. Yunchuan Sun; Houbing Song, Antonio J. Jara, Rongfang Bie "Internet of Things and Big Data Analytics for Smart and Connected Communities ", IEEE Communication Magazine,vol.4,Doi:11,10.1109/ACCESS.2016.2529723, pp. 766-773, Feb. 2016. College of Information Science and Technology, Beijing Normal University, Beijing, China
2. P. Ning and S. Jajodia "Intrusion detection techniques ", The INTERNET Encyclopedia 2003.
3. W. Feng , Q.Zhang, G. Hu and J.X. Huang, " Mining Network Data for Intrusion Detection through Combining SVMs with ant colony newtorks" Future Generation Computer Systems, vol. 37, pp.127–140, 2014.
4. D. M. Farid, M. Z. Rahman, "Anomaly network intrusion detection based on improved self-adaptive Bayesian algorithm," JournalComputing, vol. 5, no. 1, pp. 23–31, 2010
5. J. Jha and L. Ragha, "Intrusion Detection System using Support Vector Machine," International Journal Application of Information System, 2013.
6. J.F Joseph,A and Das,B.C. Seet, "Cross-Layer Detection of Sinking Behavior in Wireless Ad Hoc Networks Using SVM and FDA" IEEE Transaction on dependable and secure computing, Vol. 8, No. 2, April 2011.

7. Kyaw Thet Khaing "Enhanced Features Ranking and Selection using Recursive Feature Elimination(RFE) and k-Nearest Neighbor Algorithms in Support Vector Machine for Intrusion Detection System" International Journal of Network and Mobile Technologies 2010,1(1), 1832-6758.
8. Chie-Hong Lee, Yann-Yean Su, Yu-Chun Lin and Shie-Jue Lee "Machine learning based network intrusion detection" 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA).
9. Rohit Kumar Singh Gautam , Er. Amit Doegar "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms", 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).
10. Suad Mohammed Othman, Fadl Mutaher Ba-AlwiNabeel T. AlsohybeAmal Y. Al-Hashida "Intrusion detection model using machine learning algorithm on Big Data environment" J Big Data (2018) 5: 34. https://doi.org/10.1186/s40537-018-0145-4.
11. Kim, Kwangjo, Aminanto, Muhamad Erza, Tanuwidjaja, Harry Chandra "Network Intrusion Detection using Deep Learning "SpringerBriefs on Cyber Security Systems and Networks" 2018.
12. Chenn-Jung Huang, Ming-Chou Liu, San-Shine Chu, Chin-Lun Cheng, Application of Machine Learning Techniques to Web-Based Intelligent Learning Diagnosis System, Fourth International Conference on Hybrid Intelligent Systems (2014).
13. Ho Chun Leung, Chi Sing Leung, Eric W. M. Wong, and Shuo Li, Extreme Learning Machine for Estimating Blocking Probability of Bufferless OBS/OPS Networks, IEEE Transactions , J. OPT. COMMUN. NETW./VOL. 9, NO. 8/AUGUST 2017.
14. Fatih Ertam, Mustafa Kaya , Classification of Firewall Log Files with Multiclass Support Vector Machine , 978-1-5386-3449-3/18/$31.00 ©2018 IEEE
15. Jiexiong Tang, Chenwei Deng, Guang-Bin Huang: Extreme Learning Machine for Multilayer Perceptron, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 27, NO. 4, APRIL 2016.

AUTHORS PROFILE

Shridhar Allagi working as Assistant Professor in department of Computer Science and Engineering in KLE Institute of Technology Hubballi. His research scholar in vtu and his areas of interest are cloud security using machine learning. He has published various research articles in reputed conference and journals.

Dr Rashmi Rachh Working as Associate Professor in Dept. of Computer Science and Engineering at Visveswaraya technological University, Belgaum, Karnataka. Worked at KLE Society's College of Engineering, Belagavi. She has published more than 25 research articles in various reputed journals and conference. Her areas of research includes machine learning, cloud security and Big data Security.