A Review on Machine Learning Techniques for Intrusion Detection

Shweta Malhotra
CSE Department
Geeta Engineering College

***Abstract*— As the technology is getting advanced and the data is becoming voluminous there is a great need to retrieve only the useful data out of it. Moreover security of the data is also a big concern. There are various machine learning techniques which include data mining techniques like Support Vector Machines, Random Forests, Classification and Regression Trees, k-Nearest Neighbour Classifier, Decision Trees which are helpful in detecting thenormal data out of the abnormal one. Genetic Programming and Genetic Algorithm are other types of techniques which are utilised for recognizing only the novel features in the data. This paper surveys some of the approaches used in the literature for intrusion detection.**

Keywords*— KDD Cup 99;Intrusion Detection; Anomaly Detection,Misuse Detection; Hybrid Intrusion Detection Approaches.

I. INTRODUCTION

Network security is the biggest concern nowadays because all our computer network connections are increasing day-by-day. The term network security means to protect our networks from any suspicious activities like any unlawful access, revealing of any secret information, fabrication of data, misuse of sensitive information etc. The reasons for the sensitivity in the networks is thatan attacker can attack from any location, secondly the information is shared among the networked computers, thirdly the information has to travel through various nodes so as to reach destination and moreover each node has its own security policies and it is not mandatory that every node that receives the forwarded packet follows the same security rules.Our networks are now a target for many attackers. The attacker may be present inside the system or outside the system. The internet is the basic source for sharing of information. There are various threats to network like DoS[8], unauthorised access, where in the former attacker tries to overload the server with bulk of requests and in the latterattacker tries to access the confidential information by unauthorised ways. So, to protect ourselves from all these

unlawful events, there is a great demand for Network Security like Cryptography is applied at the application layer, to secure TCP and IP sessions we have implemented Firewalls, Honeypots, various login and passwords mechanisms, digital signatures and an alarming system which is placed inside the network protection area and system called as Intrusion Detection System (IDS)[1], [2], [3].An Intrusion Detection System (IDS) is an alarming system which inspects all the packets going through the network and gives an alert if any suspicious activity is felt by it, in the network. IDS and Firewalls both are meant for the network security. Firewalls are placed in between the inside and outside of the network and filter out the awful traffic from the sophisticated one[5]. Its only task is filtering of the bad traffic and it prevents the network from the occurrence of intrusions, whereas IDS warns the user if any suspicious activity is discovered by it[6]. The illegal packets may sometimes be passed through firewalls and IDS has the potential to detect the attack and signals an alarm to the user. IDS systems can be divided into two categories Misuse Detection and Anomaly detection. The former uses the known attack patterns where motive is to find that intruder which cracks into the system by accomplishing some known vulnerability, whereas the latter IDS Systems warns if any deviation from normal activity is encountered. According to the resources they monitor, IDS systems are categorized into two classes: Host based IDS systems and Network based IDS systems[23]. In host based the Intrusion Detection System (HIDS), scans the actions of hosts or individual computers, like the examined information is CPU time, keystroke, command sequences and system calls whereas in network based all the packets that are flowing through the network are analysedlike re-naming the content of the packet. Network Based Intrusion Detection System (NIDS) is further classified as on-line NIDS and off-line NIDS. In on-line NIDS, the data which is meant for testing for detecting whether it is intrusive or not is taken from Ethernet based connectivity, and the process for detection proceeds in real time, whereas in off-line NIDS the data is taken from some stored files, and then passed for evaluation process for testing[25].So, all the resources of the system must be protected like files, various system resources etc. against any unlawful acts.

II. RELATED WORK

Crosbie M. et al. (1995) [1] In their work they have choosen the Genetic Programming approach for detecting the intrusion. For detecting the anomaly intrusions they have used the concepts of agents and these agents are multiple in numbers. On agents a fitness score is assigned and heavy penalty is charged on those agents who misdirect the intrusions. The Automatically Defined Functions(ADF) helps in generating type- safe parse trees and every agent has multiple ADFs.Agent 2 and 3 performed better than agent 1 when three test files were provided to them.

Mukkamala S. et al. (2004)[4]. In their work they have used the DARPA 1998 dataset and the effectiveness of Genetic Programming was calculated in detecting intrusions.The performance of LGP was compared with Neural Networks and SVM and LGP outperforms in detecting intrusions.In every class the accuracy is 99% above. The performance of SVM was better than RBF and slightly less than LGP .

Wei et al. (2004)[5]. In their work they have implemented rule based approach with genetic programming. The dataset choosen by them is DARPA where 10,000 network connections were taken. The tree with a string data structure was presented i.e. "AabAcdAceI" where I means Intrusion, A means "and", and a, b, c, d represents conditions in the rule and. So, the rule is depicted as "if a and b and c and d and e, then intrusion."FPR, FNR and UADR are the three performance metrics.

Muni D.P. et al. (2004)[6].They have proposed a novel approach in designing the classifier by using Genetic Programming.The modified crossover and mutation operator were used. Directed mutation would help in not only selecting those solutions that improve the solution but also welcomed those solutions that can improve the solution.

Chebrolu S. et al.(2005)[7]. In their work, for the purpose to select only main features two algorithms were preferred namely Bayesian Networks(BN) and Classification and Regression Trees (CART). The BN used 41 variable dataset and 17 variable reduced dataset. The results show that using the latter there is improvement in performance. The ensemble of BN and CART will further help in enhancing the performance which was not possible by using them individually. For normal,probe and DOS it was 100%, for U2R it was 84% and for R2L it was 99.47%

Folino G. et al.(2005) [8].In their work, KDDCUP 1999 dataset was selected and for detecting intrusions GEdIDS (Genetic Programming Ensemble for Distributed Intrusion Detection System) was followed.The designed model was found out to be scalable, flexible and extensible. The system known as dCAGE which stands for distributed Cellular Genetic Programming System was used for executing the Genetic Programs and the algorithm used was cGp i.e. cellular GP. The task of detecting intrusions was completed by peer islands.A confusion matrix was created and the test results show that for U2R and R2L the results are worse. The GEdIDS performance is better than Linear GP.

Peddabachigari et al. (2007)[10]. In their work they have designed the hybrid systems named as Decision Trees (DT) and Support Vector Machines (SVM) that results in the formation of hybrid intelligent system which forms the hybrid intelligent system named as (DT-SVM) and an ensemble approach is taken which binds the base classifier.The dataset followed was KDDCUP1999.The results shows that accuracy achieved by ensembleapproach is 100% and for R2L and U2R 97.16 and 68% respectively. SVM works satisfactory for DOS with 99.92% accuracy.For normal class Hybrid DT-SVM showed 99.70% accuracy.

Bhavsar Y. et al. (2013)[16].In their work they have proposed a new approach in detecting intrusion by using NSL-KDDCup dataset with the SVM classifier. The preferred dataset was modified version of KDDCup dataset. So, for data pre-processing three steps are followed:

1) Data Set Transformation
2) Data Set Normalisation
3) Data Set Discretizaton

The experimental results showed the accuracy of 94.1857% and the time taken in building the model was 77.07seconds.

Dastanpour A. et al.(2013)[21].In their work, they have proposed a feature selection methodwhich is used with the GA-SVMmodel with the motive to increase the performance.FFSA and LCFS are used in detecting attacks. The dataset used was of KDDCUP 1999. The studies shows that GA with SVM and FFSA requires only 31 features to detect the attack while for Linear Coorelation feature selection(LCFS) requires 21.GA is an evolutionary process and it's main aim is to achieve global optimization by selecting only those candidates which have high fitness and eliminating low fitness candidates.The results shows that GA-SVM from feature number 21 shows 100% accuracy while FFSA from feature number 31-35 can achieve 100% accuracy.Similarly detection rate of GA-SVM is also larger than LCFS.The false positive value of GA-SVM lies in the range of 0.43%-0.6%.

Acosta-Mendoza N., et al. (2014) [20]. In their work they have suggested to use a novel approach using genetic programming for building heterogeneous ensembles. Ensemble learning is a novel approach aiming at combining various individual classifiers' output for performance improvement. The main focus of this paper is on ensemble of heterogeneous classifiers. The result shows that the method proposed in this paper is highly successful at building very effective models.

AbdElrahman et al. (2014)[19]. In their work the problem which is tackled is about class imbalance, increase detection rates for each class and minimize the false alarm in intrusion detection. In this paper a test performed on seven classifier using bagging and adaboosting ensemble methods. A new hybrid ensemble based on error Error Correcting Output Codeapproach was designed .In terms of detection rate except SVM all classifiers show satisfactory detection rate .SVM shows least detection rate for Class 1. The class 4 which has least number of samples has worse detection rate as computed by all classifiers. The new approach presented by this paper improves the accuracy (99.7%). It also increases detection rates and reduces false alarm even for the minority classes.

Dastanpour et al.(2014)[23]In their work an ensemble of GA(Genetic Algorithm) is used with ANN (Artificial Neural Network) was proposed. For retrieving only the important features Forward Feature Selection (FFS) was followed. Modified Mutual Information Feature Selection(MMIFS)uses greedy selection and hence evaluate the common features and LCFS(Linear Correlation Feature Selection) which performs classification by reducing the dimensions of the dataset. The whole process was carried out on KDDCup dataset. 100% detection is achieved by GAANN from the feature number 8, with FFSA from the feature number 31-35, with LCFS it was from feature number 21, with MMIFS it was attained in the feature number 24.

M. Govindarajan(2014)[23]. In their work the evaluation of the performance by taking homogeneous classifier named as bagging and heterogeneous classifier named as arcing was used. The choosen dataset were NSL KDDCUP and Acer07.Table 1 illustrated the accuracy of the individual and hybrid classifiers.

Parati N. et al. (2015)[25]. In their work, a hybrid technique which was followed in detecting intrusion was GA with SVM for the motive to detect intrusive activities.The performanceof hybrid RBF-

Table 1 Performance of Base and Bagged Classifier[23]

Dataset	Classifiers	Accuracy
Acer07	RBF	99.53%
	Bagged RBF	99.86%
	SVM	99.80%
	Bagged SVM	99.93%
NSL-KDD	RBF	84.74%
	Bagged RBF	86.40%
	SVM	91.81%
	Bagged SVM	93.92%

SVM classifier was better than base classifier. Whereas the bagged method was better than the base classifier. Table II shows the performance of the system.

Table II Performance of Base and Hybrid Classifier[25]

Dataset	Classifier	Accuracy
Acer07(Real Dataset)	RBF	99.40%
	SVM	99.60%
	Hybrid RBF-SVM	99.90%
NSL-KDD(Benchmark Dataset)	RBF	84.74%
	SVM	91.81%
	Hybrid RBF-SVM	98.46%

III KDD 99 DATASET

Since 1999, KDD'99 has been the most popular data set and it was specially formulated for the purpose to detect anomaly intrusion detection. This dataset was the modified version of DARPA'98. The 7 weeks of raw tcp_dump data was processed into 5 million connection records and two weeks data comprises of 2 million connection records. The whole dataset contains 41 features in which 24 types of attacks were encountered and these 24 types are further categorized into 4 types which are stated in the Table III.

TABLE III TYPES OF ATTACK

Denial of Service (DoS)	Back Ping, Smurf, Apache2
User to Root Attack (U2R)	Perl, Xtem, Load Module, Fd Format
Remote to Local (R2L)	Ftp_write, Guest, Imap, Sendmail
Probing Attack	Satan, IP Sweep, Nmap, Saint

III. METHODOLOGIES

In my research work I'll be using hybrid approach of Genetic Programming with KNN-SVM Classifier. The Dataset named asNSL-KDD [9], [10] and [19] will be used.Firstly the data is divided into two segments i.e. training data and testing data. Then after pre-processing, first feature selection is done by the GP and then it is passed to the KNN-SVMClassifier.K-NN will perform in low dimensional space where the data points are high in number and SVM will be used in that case where data points are low in number with high dimensional space. GP ensemble with KNN-SVM can be used to increase thedetection rate.

IV. CONCLUSION

After studying various literature surveys it can be concluded that for the intent to detect novel attacks Genetic Programming is the best and its benefits increases if it is used with ensemble approaches. The various classifiers like Decision Trees, SVM, Naïve Bayes, GA, Neural Networks, GA, KNN etc were used which helps in anomaly intrusion detection.Decision trees does not work efficiently in case of un- correlated data variables and for every small change in the data values a different tree is obtained. KNN classifier can be preferred in case of huge amount of data.SVM is a binary classifier and with its RBF it shows supreme results. SVM performs best in case of low number of data points with high dimensional space. If the documents number is less,then Naïve Bayes gives satisfactory results. KNN shows best results in case of voluminous features but at the same time SVM fails to perform in case of huge number of features.

V. FUTURE WORK

The research will be further carried out to determine that the GP based Ensemble Classifier is effective for reducing false alarm rate. The work can be extended by using hybrid classifier KNN-SVM-NN with Genetic programming by using NSL-KDDCup dataset.

REFERENCES

[1] Crosbie, M. & Spafford, G. ”Applying Genetic Programming to intrusion detection”. In working Notes for the AAAI Symposium on Genetic Programming , pp. 1-8. Cambridge, MA: MIT Press, November 1995.

[2] Goebel M., & Gruenwald, L. "A survey of data mining and knowledge discovery software tools." ACM SIGKDD explorations newsletter 1(1) ,pp. 20-33, 1999.

[3] Lazarevic, A., Ertöz , L., Kumar, V., Ozgur A., & Srivastava, J. "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection." In SDM, May 2003.

[4] Muni, D.P., Pal N.R., & Das,J(2004). "A novel approach to design classifiers using genetic programming."IEEE Transactions on evolutionary computation, 8(2), pp. 183-196, 2004

[5] Lu, W., & Traore, I. "Detecting new forms of network intrusion using genetic programming." Elseviere International journal of Computational Intelligence 20(3), pp. 475-494, 2004.

[6] Mukkamala S., Sung, A. H., & Abraham, A. "Modeling intrusion detection systems using linear genetic programming approach."In International Conference on Industrial, Engineering and other Applications of applied Intelligent Systems ,pp. 633-642. Springer Berlin Heidelberg, May 2004

[7] Chebrolu, S., Abraham A., & Thomas J.P."Feature deduction and ensemble design of intrusion detection systems." Elseviere International journal of Computers & Security 24(4), pp. 295-307, 2005.

[8] Folino G., Pizzuti, C., & Spezzano, G."GP ensemble for distributed intrusion detection systems." In International Conference on Pattern Recognition and Image Analysis , pp. 54-62 Springer Berlin Heidelberg. August 2005

[9] Colas, F., and Brazdil, P. "Comparison of SVM and some older classification algorithms in text classification tasks." In Artificial Intelligence in Theory and Practice. IFIP 19TH World Computer Congress, TC 12: IFIP AI 2006 Stream, August 21-24,2006, Santiago, Chile(Vol. 217, p. 169)Springer, October 2006

[10] Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. "Modeling intrusion detection system using hybrid intelligent systems." Journal of network and computer applications, 30(1)(2007),pp. 114-132.

[11] Banković,Z.,Stepanovic, D., Bojanic, S., & Nieto-Taladriz,O."Improving network security using genetic algorithm approach." Computers & Electrical Engineering,33(5), pp. 438-451,2007

[12] Münz, G., Li, S., & Carle G."Traffic anomaly detection using k-means clustering." GI/ITG Workshop MMBnet, September 2007

[13] Wu, Xi., Kumar, V., Quilan, J.R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. "Top 10 algorithms in data mining." Knowledge and information systems, 14(1) ,pp. 1-37, 2008

[14] Wang, W., Zhang, X., Gombault, S., & Knapskog, S. J."Attribute normalization in network intrusion detection." In 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks , pp. 448-453 IEEE, December 2009.

[15] Modi, C., Patel D., Borisaniya, B., Patel, A., & Rajarajan, M. ."A survey of intrusion detection techniques in cloud." Journal of Network and Computer Applications 36(1), pp. 42-57, 2013

[16] Bhavsar, Y.B., & Waghmare, K.C. "Intrusion detection system using data mining technique: Support vector machine." International Journal of Emerging Technology and Advanced Engineering, 3(3) ,pp. 581-586,2013

[17] Zargar, S.T., Joshi J., & Tipper, D."A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks." IEEE Communications Surveys & Tutorials 15(4) ,pp.2046-2069.

[18] Dastanpour, A., & Mahmood, R.A.R.."Feature selection based on genetic algorithm and Support Vector machine for intrusion detection system." The Second International Conference on Informatics Engineering & Information Science (ICIEIS2013) (pp. 169-181). The Society of Digital Information and Wireless Communication, 2013.