IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

A Comprehensive Analysis of Machine Learning Models for Real Time Anomaly Detection in Internet of Things

S.L. Sanjith[1*], E George Dharma Prakash Raj[2]

[1] Indian Institute of Management Tiruchirappalli, Tamilnadu, India,
[2] Bharathidasan University, Tiruchirappalli, Tamilnadu, India

[*]*Corresponding Author: sanjithsl@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 22/Nov/2018, Published: 30/Nov/2018

***Abstract*—** Anomaly detection is a major requirement of the current Internet of Things (IoT) and inter-networked communication environment. This work analyzes recent and prominent contributions in the domain of Anomaly detection. The analysis is performed especially in domains related to real time operations and IoT environment. The review is performed and results from most prominent models in literature are considered for analysis. This paper discusses the working mechanisms and the major issues in Anomaly detection such as data imbalance and noise especially in IoT domain and the methods used to handle these issues. Experiments were performed using the NSL-KDD benchmark data set. Precision, False Positive Rate and Accuracy are used to analyze the effectiveness of the models.

***Keywords*—** Multi-Layered Clustering, Ensemble Models, Intrusion Detection, K-Means, SVM

I. INTRODUCTION

The Internet of Things (IOT) era has made a huge rise in the number of devices. Further, the capabilities of the devices have also seen a huge boost, resulting in increased capabilities of these devices to exchange information [1]. The increased availability of such devices and the huge reduction in cost levels and the cost of operations has resulted in increased adoptions of such technologies and devices. This has resulted in large number of users utilizing such technologies. The increase in usage levels has resulted in a large number of information flow within the network [2]. This has inevitably attracted hackers. The emergence of Big Data techniques and the reducing network boundaries has resulted in increased network Anomalies. Traditionally, anomaly detection and prevention was performed using Firewalls, user authentication mechanisms and data encryption methods. However, currently, such methods are limited in capabilities and cannot handle huge amount of data. This mandates new technologies for detecting and preventing network anomalies [3].

anomaly Detection System is the process of analysing the network traffic and identifying if an anomaly has occurred. These models are mainly categorised into two; anomaly-based systems and misuse-based systems [4, 5]. Misuse based models are classifiers built using anomalous data for training the model. It has its major focus towards detecting anomalous traffic. The rules are determined using anomalous signatures. Any network data matching these signatures are flagged by these models. The major downside of these models is that anomalous traffic is dynamic and varies over time. Hence models trained with past signatures cannot cope with variations in the traffic. Anomaly based systems are built using normal traffic as their base data. Data showing variations from the normal instances are classified as anomalous. However, it should be noted that normal data is huge and hence the trained model should be capable of utilizing a large amount of data for model building phase.

Machine learning algorithms are widely used for anomaly detection. Both supervised and unsupervised models can be effectively used for the prediction process [6]. Further, with the increase in the complexity levels of anomalies, a single algorithm might not be suitable for complex predictions. Hence the current literatures witness a large usage levels of ensemble-based models for Anomaly detection. Ensembles are models built using multiple and sometimes varied models for prediction. Results from all the models are aggregated to obtain the final prediction. Such models are widely used due to the increasing complexities and reducing distinctions between normal and anomalous traffic [7,8].

This work deals with analysis of recent and most prominent machine learning algorithms used for Anomaly detection. The work analyzes Semi-Supervised Multi-Layered Clustering (SMLC) model, RampLoss K-SVCR model, LMDRT-SVM and LMDRT-SVM2 and Extreme Learning Machine (ELM) based prediction models. Analysis is performed based on Precision, False Positive Rate and Accuracy.

The rest of this paper is organized as given below; section II presents issues and challenges in IDS, section III presents the working of existing machine learning models, section IV presents the description of the Anomaly detection data (NSL-KDD) used for analysis, section V presents an study of the results and section VI determines the work.

II. ANOMALY DETECTION : ISSUES AND CHALLENGES

Anomaly detection system can be formulated as binary classification or multi-class classification model. The multi-class classification model considers normal traffic as a class and each type of anomalous traffic is indicated as a new class, while binary classification model considers normal traffic as the negative class and all the other classes as positive class. Both considerations have their own pros and cons and are used depending on the result requirements. The domain however has several intrinsic properties associated with it, which also poses challenges, irrespective of the type of classification used. They are, data imbalance, noise, outliers and data hugeness.

A. Data Imbalance

Data imbalance is the property of a dataset to contain a large number of instances pertaining to one class and very low instances pertaining to another class. Network generated data is laden with a large number of normal instances, while the anomalous or attack traffic tends to be very low, hence indicating the presence of imbalance [9]. The major issue in using a data with data imbalance is that the imbalance levels tends to create a biased classifier model, where the model is over trained with the majority class and undertrained with the minority class [10].

B. Presence of Noise and Outliers

The network data, being machine generated, records all the generated packets, leading to noise and outliers [11,12]. The presence of noise and outlies tend to heavily affect the generalization capabilities of the classifier models [13]. Noise is considered to be a variation in the regular data, while outliers are exaggerated noise levels exhibiting very high variations, positioning themselves as belonging to a different class. This usually occurs due to variations in the distribution levels of the data. Classifiers, when operated upon noisy data and outliers, interprets them as a different class. However, when the training data is labelled, such instances disrupt the decision rules, reducing the prediction efficiency of classifier models [14].

C. Big Data

The increase in the usage of interconnected devices has caused the generation of enormous amounts of information. Processing such huge amounts of information requires Big Data environments for effective processing. However, large number of instances alone is not the major issue, instead an increase in the dimensionality levels of the data was also observed in the current network data. This leads to the curse of dimensionality, where the learning models get computationally intensive, which successively influence the performance of the classifier models.

III. ANOMALY DETECTION MODELS: AN ANALYSIS

Anomaly detection has been under research since the inception of networks and communication technologies. The increased usage of such technologies has resulted in several research articles dealing with anomaly detection and prevention systems. Although being a legacy topic, it is still a hot topic of research due to the ever-changing Anomaly signatures and the reducing distinction between the normal and anomalous traffic in the network. This work considers four recent, distinct and most prominent models for analysis; the ensemble based Semi-Supervised Multi-Layered Clustering (SMLC) [15] model, SVM based model Ramp K-SVCR [16], Feature Augmentation based models LMDRT SVM and LMDRT SVM2 [17] and Extreme Learning Machine (ELM) based Clustering model [18].

A. Semi-Supervised Multi-Layered Clustering (SMLC)

SMLC is an ensemble-based model that performs semi-supervised learning to perform Anomaly detection. The SMLC model is a combination of supervised and un-supervised learning methods. This model is composed of two major phases; partially labelled data points are initially clustered to identify their corresponding classes, which is followed by Decision Tree based supervised learning to detect anomalies.

K-Means clustering algorithm is used for clustering process. This operates by iteratively grouping *K* clusters until a satisfactory grouping is obtained. The grouping is validated using Euclidean Distance as the base measure. The proposed model modifies this by proposing a weighted Euclidean Distance measure for clustering. The groups identified mainly depends on the initial centroid points selected for the grouping process. This model generates multiple layers of clustered instances and finally combines them to provide the final grouped clusters. Overlaps in multiple layers are used as the base for generating the final groups. The groups formed is used to provide labels for the unlabeled instances in the training data. The final labelled instances are used to train the Decision Tree model for prediction.

The main benefit of this model is that it performs the modelling on partially labelled data, hence effectively handling the issue of varying distributions effectively. The

main disadvantage of this model is that it is computationally intensive, as the model requires multiple training algorithms for a single prediction.

B. Ramp K-SVCR

The Ramp K-SVCR is an SVM based model, aimed to specifically handle data imbalance and the skewed distributions of the network data to perform effective anomaly detection. The use of K-Support Vector Classification-Regression (K-SVCR) is for performing multiclass classification, while the Ramp function is used to handle noise and outliers effectively.

The Ramp K-SVCR has been specifically developed to handle multiclass classifications in network data. The K-SVCR is a multiclass approach that operates based on 1-vs.-1-vs.-Rest model. The existing K-SVCR model uses Hinge loss function, leading to the model getting highly sensitive towards noise and outliers. Hence the Ramp K-SVCR replaces it with the Ramp loss function, which is highly tolerant towards noise. It uses Alternating Direction Method of Multipliers (ADMM) procedure to reduce the training time and perform prediction on large-scale data, and the Concave-Convex Procedure (CCP) to resolve the model.

The key benefits of this model is that it is generalizable and could be used on data with imbalance and noise levels. This improves the usability of the proposed model to a large extent. Downsides of this models is that the model was unable to handle data with very high imbalance levels. Further, it can be perceived that the nature of the proposed model is static, hence affecting the generalizability of the model.

C. LMDRT SVM and LMDRT SVM2

Feature Augmentation is the process of identifying or generating best features from the given data set, so as to provide quality data to the machine learning model. This model uses LMDRT for feature augmentation and SVM for prediction.

The Logarithm Marginal Density Ratios Transformation (LMDRT) is a process to generate more prominent features from the existing data. The process of LMDRT works by initially splitting the data into two distinct partitions. The kernel density estimation is applied on the data and data transformation is performed. LMDRT results in making the underlying class differences between features becoming more prominent. The generated transformed data is passed to SVM for creating an Anomaly detection model. A variant of LMDRT-SVM, the LMDRT-SVM2 has also been proposed in this work. The LMDRT-SVM2 exhibits a slight variation, where both the transformed and the un-transformed data is used for the SVM training process, whereas only the transformed data is used for the training process in LMDRT-SVM.

The major advantage of this model is that LMDRT is a effective feature transduction technique that can result in high quality data. Feature Augmentation further serves to reduce the training time, leading to a faster model. Main disadvantage of this model is that SVM being a legacy model has two major constraints of being computationally complex and being prone to data imbalance. As both these issues are constituents of the domain, the generalizability of SVM is reduced to a very large extent.

D. Extreme Learning Machines (ELM)

The clustering based Extreme Learning Machine (ELM) model is built to enable effective Anomaly detection and expert interaction to fine-tune the
prediction process. Apart from the prediction module, the Anomaly detection architecture is composed of three major components; the Clustering Manager, Decision Maker and the Update Manager.

The cluster manager constructs the machine learning model by mapping the training data into clusters. Each obtained cluster is considered to represent a class. The base system is constructed using Extreme Learning Machines (ELM). This work uses a modified version of CLUS-ELM, a clustering based extreme learning machine for prediction. These predictions are not directly utilized, instead, they are passed to the Decision Maker for analysis. The Decision Maker analyzes the predictions, identifies errors and provides correction proposals to the human expert. The human expert analyzes these suggestions and identifies further modifications that are required for the model and provides it to the update manager. The update manager updates the clustering model by fine-tuning it towards the provided suggestions.

The key benefit of this model is that it also involves a human expert. This leads to the model becoming flexible and fine-tunable towards changes or variations contained in the domain data. The major downsides of ELM is that it is complex in terms of computation and hence might not be capable of handling huge real-time data.

Table 1. Features of the NSL-KDD dataset

Class	Feature	Data Type
Basic features	duration	continuous
	protocol type	nominal
	service	nominal
	src_bytes	continuous

	dst_bytes	continuous
	flag	nominal
	land	nominal
	wrong _fragment	continuous
	urgent	continuous
Content – based features	hot	continuous
	num_failed_logins	continuous
	logged_in	nominal
	num_compromised	continuous
	root_shell	nominal
	su_attempted	nominal
	num_root	continuous
	num_file_creation	continuous
	num_shells	continuous
	num_access_file	continuous
	num_outbound_cmds	continuous
	is_hot_login	nominal
	is_guest_login	nominal
Time – based traffic features	count	continuous
	serror_rate	continuous
	rerror_rate	continuous
	same_srv_rate	continuous
	diff_srv_rate	continuous
	srv_count	continuous
	srv_serror_rate	continuous
	srv_ rerror_rate	continuous
	srv_diff_host_rate	continuous
Host – based traffic features	dst_host_count	continuous
	dst_host_srv_count	continuous
	dst_host_same_srv_rate	continuous
	dst_host_diff_srv_rate	continuous
	dst_host_same_src_port_rate	continuous
	dst_host_srv_diff_host_rate	continuous
	dst_host_serror_rate	continuous
	dst_host_srv_serror_rate	continuous
	dst_host_rerror_rate	continuous
	dst_host_srv_rerror_rate	continuous

IV. DATASET DESCRIPTION

Analysis is performed using the NSL-KDD data set. The NSL-KDD dataset was derived from the KDD CUP 99 dataset [19], which is a standard dataset for Anomaly detection. KDD CUP 99 dataset however has several issues like duplicated records, leading to bias in the training process of classifier. NSL-KDD dataset is constructed by eliminating all the duplicates to provide a cleaner and compact dataset [20]. Table 1 shows the features of the NSL-KDD dataset.
NSL-KDD is a multi-class classification dataset, composed of five classes; normal and four attack classes. Normal class instances are the majority class instances with highest frequency, while all the others are minority class instances, with U2R exhibiting very high imbalance levels. The detailed descriptions are shown in table 2.

Table 2. Details of the NSL – KDD data set

Class	Traffic - type	Full NSL – KDD training set	
		Number of records	Frequency
1	Normal	67,341	53.46
2	DOS	45,927	36.43
3	Probe	11,656	9.25
4	R2L	995	0.79
5	U2R	52	0.04
Total records		125,971	

V. PERFORMANCE ANALYSIS AND DISCUSSION

The NSL-KDD dataset has been applied to the existing models and the obtained results are analyzed to identify the best performing model. Analysis has been performed based on three major metrics; Precision, False Positive Rate (FPR) and Accuracy.

Figure 1 shows the comparison of the precision values. It could be observed that the best performances were exhibited by SMLC and LMDRT-SVM and LMDRT-SVM2, followed with a very slight difference by the RampLoss K-SVCR. ELM was observed to exhibit reduced precision levels at 0.84.

A comparison of the False Positive Rates (FPR) is shown in figure 2. Lower FPR levels indicate better models. It could be observed that the LMDRT SVM and LMDRT SVM2 exhibits very high FPR levels indicating the presence of a huge false alarm rate. The other models RampLoss K-SVCR, SMLC and ELM exhibits low FPR levels showing desirable conditions for an Anomaly detection system.

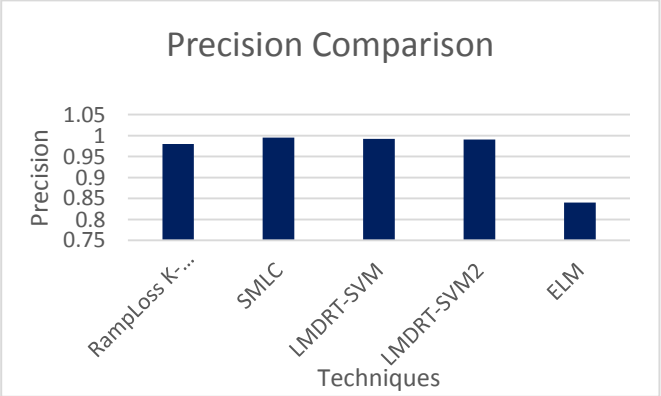

Figure 1. Comparison of Precision

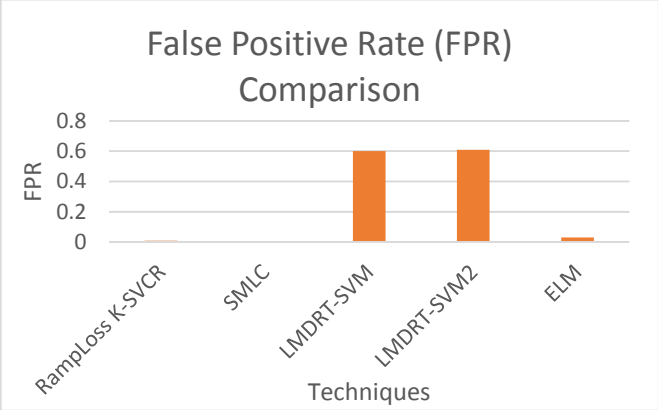

Figure 2. Comparison of False Positive Rate (FPR)

A comparison of the aggregated metrics is shown in figure 3. Accuracy is an overall metric that exhibits the predictability level of a system as a whole. It can be detected that the models, except for ELM model shows very good accuracy levels.

Table 3 shows a tabularized representation of the results. The best results are shown in bold. An overall analysis indicates that RampLoss K-SVCR and SMLC models exhibits the best prediction levels and hence satisfies the requirements of an IDS.

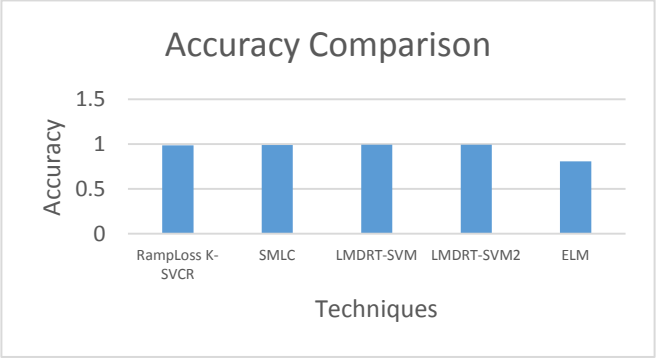

Figure 3. Comparison of Accuracy

Table 3. Performance Comparison

	Accuracy	FPR	Precision
RampLoss K-SVCR	**0.986**	**0.008**	**0.98**
SMLC	**0.99**	**0.003**	**0.995**
LMDRT-SVM	**0.993**	0.6	**0.992**
LMDRT-SVM2	**0.992**	0.61	**0.991**
ELM	0.808	0.03	0.84

VI. CONCLUSION

This work presents an analysis of the existing and prominent anomaly detection models in literature. This work uses Semi-Supervised Multi-Layered Clustering (SMLC) model, RampLoss K-SVCR model, LMDRT-SVM and LMDRT-SVM2 and Extreme Learning Machine (ELM) based model for analysis. The paper begins by addressing the major issues existing in detection of Anomalys in the IoT domain, followed by providing a detailed working mechanism of the models to be analyzed. A data description of the NSL-KDD dataset is provided to highlight its significance. The results obtained from various models on NSL-KDD dataset is analyzed in terms of Precision, FPR and Accuracy. Analysis indicates that the RampLoss K-SVCR and SMLC models exhibit effective performances and also exhibits the generalizability levels of the models.

REFERENCES

[1] N. Mohamudally, M.Mahejabeen Peermamode, "Building an Anomaly Detection Engine (ADE) For IoT Smart Applications". Procedia computer science, Vol. 134, pp.10-17, 2018

[2] S. Ahmad, L. Alexander, P. Scott, A. Zuha, "Unsupervised real-time anomaly detection for streaming data", Neurocomputing, Vol. 262, pp.134-147, 2017

[3] Mahdavinejad, S. Mohammad, R. Mohammadreza, B. Mohammadamin, A. Peyman, B. Payam, P. Sheth, "Machine learning for Internet of Things data analysis: A survey", Digital Communications and Networks, 2017.

[4] Hoque, Mohammad Sazzadul, Md Mukit, Md Bikas, and Abu Naser. "An implementation of intrusion detection system using genetic algorithm." arXiv preprint arXiv:1204.1336, 2012.

[5] Piyush Pareta, Manish Rai, Mohit Gangwar, "An Integrated approach for effective Intrusion Detection with Elasticsearch", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.13-17, 2018

[6] O.Y. Al-Jarrah, P.D. Yoo, S. Muhaidat, G.K. Karagiannidis, K. Taha, "Efficient machine learning for big data: a review", Big Data Res. big Data, Analytics, and High-Performance Computing, Vol.2, Issue.3, pp.87–93 https://doi.org/10.1016/j.bdr.2015.04.001, 2015

[7] S. Abt, H. Baier, "A plea for utilising synthetic data when performing machine learning based cyber-security experiments", in: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, ACM, pp.37–45, 2014.

[8] Ramchandar Durgam and R.V.Krishnaiah, "Online Intrusion Alert Aggregation with Generative Data Stream Modeling", International Journal of Scientific Research in Computer Science and Engineering, Vol.1, Issue.5, pp.23-23, 2013

[9] S.M.H. Bamakan , H. Wang , T. Yingjie , Y. Shi , "An effective intrusion detec- tion framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization", Neurocomputing, Vol.199, pp.90–102, 2016.

[10] S. Akila, and U.S. Reddy. "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data". Proceedings of ICRECT, Vol.16, pp.28-34, 2016.

[11] J. McHugh , "Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory", ACM Trans. Inf. Syst. Secur. Vol.3, Issue. 4, pp.262–294, 2000 .

[12] P. Rutravigneshwaran, "A Study of Intrusion Detection System using Efficient Data Mining Techniques", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.6, pp.5-8, 2017
[13] J.P. Brooks , "Support vector machines with the ramp loss and the hard margin loss", Operat. Res. Vol.59, Issue.2, pp.467–479, 2011.
[14] A. Somasundaram, and U.S. Reddy. "Modelling a stable classifier for handling large scale data with noise and imbalance". In Computational Intelligence in Data Science (ICCIDS), IEEE International Conference, pp. 1-6, 2017.
[15] Al-Jarrah, O.Y., Al-Hammdi, Y., Yoo, P.D., Muhaidat, S. and Al-Qutayri, M. "Semi-supervised Multi-Layered Clustering Model for Intrusion Detection". Digital Communications and Networks. 2017
[16] S.M.H. Bamakan, H. Wang, and Y. Shi. "Ramp loss K-Support Vector Classification-Regression; a robust and sparse multi-class approach to the intrusion detection problem". Knowledge-Based Systems, Vol.126, pp.113-126, 2017.
[17] H. Wang, J. Gu, and S. Wang. "An effective intrusion detection framework based on SVM with feature augmentation". Knowledge-Based Systems, Vol.136, pp.130-139, 2017.
[18] S. Roshan, Y. Miche, A. Akusok, and A. Lendasse. "Adaptive and online network intrusion detection system using clustering and Extreme Learning Machines". Journal of the Franklin Institute, Vol.355, Issue.4, pp.1752-1779, 2018.
[19] KDD Cup'99 intrusion detection data set, Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, October 2007.
[20] M. Tavallaee, E. Bagheri, W. Lu, and A.A. Ghorbani,. "A detailed analysis of the KDD CUP 99 data set". In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium, pp. 1-6, July, 2009

Authors Profile

Mr. Sanjith S L has completed his M-Tech from Manonmanium Sundaranar University, Tirunelveli in Computer and Information Technology in the year 2012 and B-Tech in Electronics Engineering from Cochin University of Science and Technology in the year 1999. He has more than 18 years of experience in the Planning, Designing, Implementation and Management of ICT Infrastructure of reputed organizations out of which 13 + years in academic organizations. Currently he is working as Systems Administrator at Indian Institute of Management Tiruchirappalli (IIM Trichy). Before joining IIM Trichy, he was working as Senior System manager at P A Aziz College of Engineering and Technology, Trivandrum. He also worked as part-time consultant in few organizations for the implementation of ERP. His experience includes Managed LAN (OFC & Ethernet), Servers with VMs, ERP implementation, Controller based WiFi network, IP Telephony system, IP Surveillance system and Automated Audio visual solutions. This experience became the motivation for pursuing his PhD in 'Decentralized real-time security mechanism in Internet of Things' from Bharathidasan University, Tiruchirappalli.

Dr. E. George Dharma Prakash Raj completed his Master's Degree in Computer Science and Masters of Philosophy in Computer Science in the years 1990 and 1998. He has also completed his Doctorate in Computer Science in the year 2008. He has around twenty-seven years of Academic experience and nineteen years of Research experience in the field of Computer Science. Currently he is working as a Faculty in the School of Computer Science, Engineering and Applications at Bharathidasan University, Trichy, India. He has published several papers in International Journals and Conferences related to Computer Science and has been an Editorial Board Member, Reviewer and International Programme Committee Member in many International Journals and Conferences. He has convened many National and International Conferences related to Computer Science.

