Comparison Deep Learning Method to Traditional Methods Using for Network Intrusion Detection

Bo Dong
Computing Center of Liaoning University
Shenyang, China
e-mail: dongbo@lnu.edu.cn

Xue Wang*
Information Technology Center of Liaoning University
Shenyang, China
e-mail: wangxue@lnu.edu.cn

Abstract—**Recently, deep learning has gained prominence due to the potential it portends for machine learning. For this reason, deep learning techniques have been applied in many fields, such as recognizing some kinds of patterns or classification. Intrusion detection analyses got data from monitoring security events to get situation assessment of network. Lots of traditional machine learning method has been put forward to intrusion detection, but it is necessary to improvement the detection performance and accuracy. This paper discusses different methods which were used to classify network traffic. We decided to use different methods on open data set and did experiment with these methods to find out a best way to intrusion detection.**

Keywords-deep learning; intrusion detection; network security; classifier machine learning

I. INTRODUCTION

Deep learning has in recent times gained prominence due to the potential it portends for machine learning. Of importance, the science has become an integral part of network security since it ensures exhaustive and conclusive evaluation of the network security system. To start with, deep learning can be defined as the use of deep networks that are linked to calculate algorithms that in turn use several layers to produce an output. In essence, the layers cascade with the next tier using the results from the previous phase as input in order for it to produce an output.

Notably, the layers normally include the input tier that has basic data, which then is analyzed by the multiple hidden levels, and finally the last phase that produces the output. Primarily, the model relies on unsupervised features that form higher representations of information from lower levels. Further, new techniques in deep learning are being developed in response to increased data volumes and need for more accurate and conclusive evaluation. New systems include deep belief and deep coding approaches in machine learning. We give comparison to different methods in this paper. And it is indispensable to measure which is the best method for network traffic recognition and intrusion detection.

II. DEEP LEARNING IN NETWORK SECURITY

Network security is an important component in the information technology field since it provides preventive security strategies to the physical and software infra-structure of the system. Notably, the amount of data being generated by networks is on the increase within the information technology industry, which includes network machines, operating systems, and applications. On the downside however, the sector still relies on manual approaches to prevent and mitigate any malfunctions on the system.

The structure of deep learning is described in Fig. 1. Deep learning methods which are inspired by the structure depth of human brain learn from lower level characteristic to higher levels concept. It is because of abstraction from multiple levels, DBN [1] helps to learn functions which are mapping from input to the output. The process of learning does not dependent on human-crafted features. DBN uses an unsupervised learning algorithm, a Restricted Boltzmann Machine (RBM) for each layer.

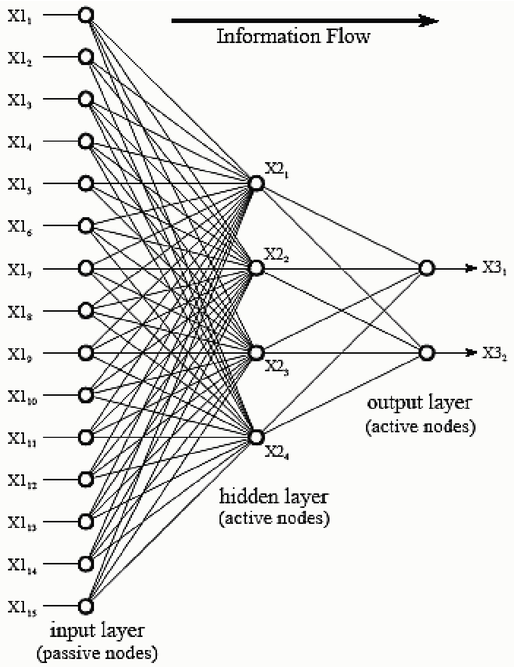

Figure 1. Structure of deep learning

A. *Traffic Identification*

Traffic identification is a key component in network security since it raises the red flag in case, of intrusion into the network. Notably, the system has relied on traditional methods of detection that are increasingly becoming

978-1-5090-1781-2/16/ $31.00 ©2016 IEEE

ineffective due to the commensurate increase in data. Traditional approaches include port identification for instance, standard HTTP that is failing to perform as envisaged due to less protocols following the system. Another system involves the signature-based method that relies on the payload data. Of importance, the approach can be used in several applications.

Researchers have put forward a lot of data mining algorithms to traffic identification. Traditional data mining methods, such as Native Bayes [2], Random Forests [3], Decision Tree [4], were widely used for classifying network traffic. Jun *et al.* [5] used SVM and RBM together to solve the problem of network traffic detection.

Nonetheless, machine learning has led to improved methods of analyzing problems in network security. Practically, deep learning can be used in analyzing network security through a number of ways. To start with, the method has proved useful in detecting any anomalies on the network system since it uses statistical data to calculate any network problems by evaluating the interrelation of neutrals in the system. In order to understand anomalies in network security it is important to assess several variables in network security.

B. Factors in Network Security

First, in trying to detect intrusion into the system there is always the challenge of differentiating between normal and abnormal data on the system. To achieve this, the detection approach should define the characteristics of malicious data on the system. Further, the technique should be able to design a classification system that is able to differentiate between the two sets of information accurately that is genuine and malicious data. This system is known as dimensionality reduction technique that uses automatic encoding approaches to calculate the distance between nodes within the network. Importantly, the technique works on the assumption that the normality of the data is determined by consistency of distance between the nodes. As such, the longer the distance between the nodes is indicative of the abnormality of the information thus acting as a pointer to the presence of malicious data. In relation to this, two measurement systems exist that are Manhattan distance, which is the total distance between the dimensions within the network and the Euclidean distance that primarily is the size of the vector being assessed.

Second, another challenge in deep learning is the sanctity of data used to detect an anomaly, in a process known as normality poisoning. Fundamentally, the features used to extract data are important since they determine the output especially in unsupervised deep learning. As such, the approach should ensure that normal information is not affected; this is to ensure that anomalous data is not hidden within normal information in the network thus making the whole process self-defeatist. Notably, the network can be manipulated using different methods that include increased traffic, and manipulation of information.

C. Techniques Used to Detect Anomalies in Network Security

Most importantly, several approaches can be used to detect anomalies in network security using deep learning method. To start with, auto encoder dimensionality reduction is a system that relies on encoder and decoder components that further include input, output, and hidden layers. Additionally, the system uses three steps that include pre training, unrolling and a system for fine-tuning.

Equally important, is the Deep Belief Network (DBN) a deep learning method that uses an unsupervised layer of RBM and another supervised tier of BP networks. The DBN method is divided into two that is the RBM is conditioned separately in an unsupervised manner. Further, the second process involves the BP neural process involves use of last output from the RBM as the input of the new BP tier, which are then classified using a supervised approach.

Consequently, the two techniques are combined to form the hybrid malicious detection system. The approach relies on the reduction of dimensionality through use of auto encoder method to provide the space between the vectors. Consequently, this data is classified through use of the DBN system through deep learning. Finally, the accuracy for detection is improved in addition, to reducing complexities associated with time in the hybrid system.

Unknown Protocol identification is a big challenge to traditional methods of detection in network security since research has established that nearly 17 percent of traffic flows on networks are unknown. However, deep learning has tried to lessen the problem since it is able to identify more than half of the unknown traffic flow in traditional methods. Better still the approach is able to place a probability on the unknown flows thus increasing accuracy.

III. Deep Learning through Analysis of Data Patterns in Network Security

The growth of the information technology field has necessitated the need for newer and better methods of analyzing how these computer systems operate. To that end, several methods in machine learning exist that try investigate the principles behind the devices. Notably, the field of deep learning is dynamic due to the development of new techniques in several sub branches that include image recognition, computer security, and speech recognition. The classical methods of deep learning used in network security are increasingly failing to detect intrusions into network systems due to the commensurate increase in data production. As such, big data analysis using deep belief system is the latest innovation that tries to study information patterns with a view of detecting unauthorized entry into computer networks.

A. Deterministic Analysis

In essence, the system operates on known principles that include deterministic and probabilistic machine learning. Deterministic machine learning uses small data sets, which are analyzed for any deviations from any normal patterns. This information is then evaluated by IT experts who formulate models to be used for further investigations of the data. Normally, the information obtained is compared against a baseline, as such any data that exceeds the normal levels is considered as intrusive action.

B. *Probabilistic Analysis*

On the other hand, probabilistic machine learning goes a step further since it evaluates the patterns involved in the assessment that might have escaped the deterministic analysis. Importantly, the system relies on clustering in order to detect any anomalous character concerning data. The system relies on unsupervised action where the system runs independently to generate a map that is eventually analyzed by the same machine for any abnormal behavior. Consequently, the approach is more efficient since the assessment is conclusive and can pinpoint the exact problem with conservative estimates placing it at 90 percent.

C. *Deep Coding Networks*

Deep coding networks have also gained traction in recent times due to the advantages it possesses in deep learning techniques. The approach is dynamic since the system is predictive and adapts itself to new data environments. Notably, the method uses outputs from top-down approaches and uses them as inputs for bottom-up approaches. In addition, the model extracts features using linear models that are in turn used as construction elements for layers that are dependent on each other that are previous and succeeding tiers to form deep system architecture.

IV. Problems of the Intrusion Detection

A. *Data Set for Evaluation*

The main problem of intrusion detection is to assure a safety communication in networks with different multi-nodes from possible network intruders by the method of classifying incoming traffic into normal and different anomalous classes. The existing intrusion detection methods sustain from false positive detection or negative detection problem due to the reason of that a lot of intrusion actions remain undetected and legitimate users are detected as intruder. In this paper, we make comparison of accuracy with different methods on the KDD-99 [6] data-set.

Since 1999 KDD-99 has been the most wildly used data set for the evaluation of traffic classify methods. The attacks in KDD-99 data set fall in four categories in [6] as following:

1) *Denial of Service Attack (DOS):* is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

2) *User to Root Attack (U2R):* is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

3) *Remote to Local Attack (R2L):* occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

4) *Probing attack:* is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

B. *Data Set Pretreatments*

Due to the imbalance of the KDD-99 data set in Table I, all the methods to classify the traffic in it become difficulty.

TABLE I. Distribution of Attacks in KDD Cup 99 IDS Dataset

Title of Dataset	**Data Classified**					
	Normal	***DOS***	***Probe***	***U2R***	***R2L***	***Total***
10% KDD Data	97278	391458	4107	52	1126	494021
10% KDD Data For Test	60591	223298	2377	39	5993	292298

To solve this problem, Synthetic Minority Over-Sampling Technique (SMOTE) which is a very popular technique has been applied to deal with this issue [7].

V. Performance Evaluation

After SMOTE technique procedure, the imbalance problem has been solved. The results are presented in Table II.

TABLE II. Distribution of Attacks after Applied SMOTE

Title of Dataset	**Data Classified**					
	Normal	***DOS***	***Probe***	***U2R***	***R2L***	***Total***
10% KDD Corrected Data	559186	391458	726993	671372	735472	3084481

The imbalance distribution problem has been reduced in Table II. Then we use the data to our experiments. Some of the features are listed in Table III.

TABLE III. Some of Features Used in Experiments

No.	***Name***	***Type***
1	HTTP response code	Number
2	HTTP request type	Text
3	HTTP packet length	Number
4	Contain attachment	Number
5	Attachment type	Text
6	Attachment size	Number
7	Download/upload	Boolean
8	Number of HTTP links with same IMSI in 2 min	Number
9	Number of HTTP packet sent with same IMSI in 2 min	Number
10	Number of HTTP packet received with same IMSI in 2 min	Float
11	Send-to-receive ratio of packets with same IMSI in 2 min	Float
12	Number of bytes sent with same IMSI in 2 min	Number
13	Number of bytes received with same IMSI in 2 min	Number
14	Send-to-receive ratio of bytes with same IMSI in 2 min	Float
15	Ratio of packet with the same destination IP in 3 min	Float

It is a classification issue for traffic detection. We select some classic classification methods, such as Decision Tree (C4.5) [8], Naïve Bayes [9], Support Vector Machine (SVM) [10], and SVM-RBMS [5].

VI. COMPARISON OF TRADITIONAL AND NEW METHOD OF DEEP LEARNING

A. *Evaluation Indicators*

Widely used indicators are *Precision* as in (1) and Recall as in (2).

$$Precision=TP/(TP+FP) \tag{1}$$

$$Recall=TP(TP+FN) \tag{2}$$

where:

TP: corresponds to the number of positive examples correctly predicted by classifier.

FN: corresponds to the number of positive examples wrongly predicted as negative by classifier.

FP: corresponds to the number of negative examples wrongly predicted as positive by classifier.

TN: corresponds to the number of negative examples correctly predicted as negative by classifier.

B. *Hardware and Software Configuration*

We choose same hardware and software configurations when carrying out the experiments. It can be found the main items of our hardware and software configuration in Table IV.

TABLE IV. HARDWARE AND SOFTWARE CONFIGURATION

No.	***Hardware or software***	***Type***
1	Operating system	Ubuntu 14.04.4 LTS
2	Programming language	Java with JDK1.7
3	File system	HDFS
4	Hadoop version	1.2.1
5	Spark version	1.3
6	Cluster	8
7	CPU	Xeon E5-46032.00GH
8	Core number of CPU	8
9	RAM	128GB DDR3
10	Disk	8TB with RAID 5

C. *Results*

To found out the best performance, we use different methods to evaluate which algorithm could be the best precision. We choose precision as the main evaluation standard. The results can be seen in Fig. 2-Fig. 6. These figures show the precision when we use different algorithm on different traffic. The best result is appeared when we use SVM-RBMs learning method to classify normal traffic in Fig. 2. It is due to the normal traffic is big enough to be recognized. Along with the percent of training samples increased, the precision of each algorithm gets better. And the method of SVM-RBMs gets the better performance than others.

We find the fact that SVM-RBMs get the best precision. SVM and C4.5 have the nearly the same precision when classifying normal traffic. All the algorithms have lower precision when they classify UToR and RToL traffic. The reason of that is that the amount of UToR and RToL packages is too lower than normal packages what we have mentioned above. But the precision to classify UToR traffic has been greatly improved by introduced SMOTE. The result can be seen in Fig. 7. The precision gets much better when we classify UToR traffic with SMOTE, though it cannot be an ideal value, than using SVM-RBMs without SMOTE.

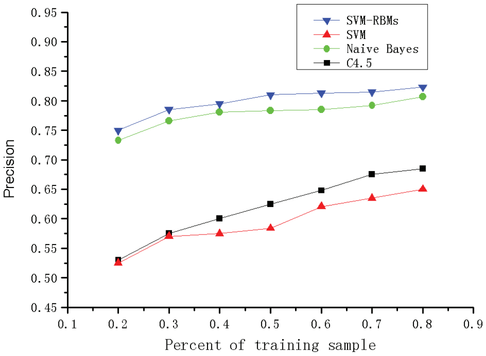

Figure 2. Precision of different algorithms used to recognize normal traffic

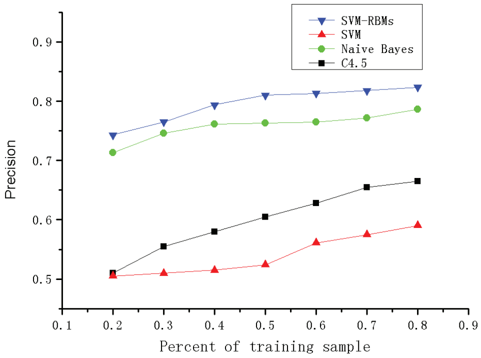

Figure 3. Precision of different algorithms used to recognize DOS traffic

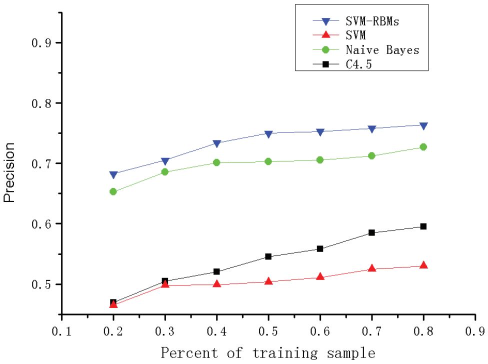

Figure 4. Precision of different algorithms used to recognize DOS traffic

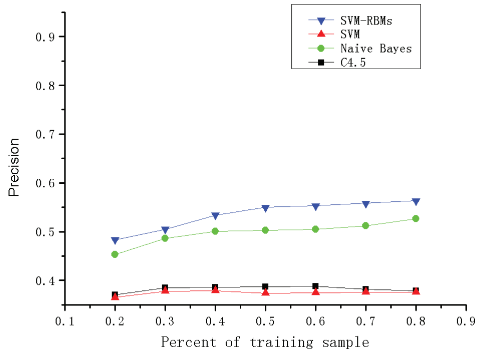

Figure 5. Precision of different algorithms used to recognize UToR traffic

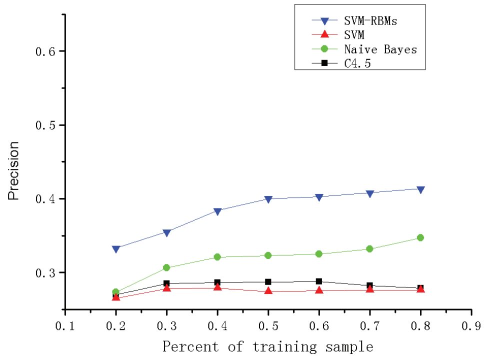

Figure 6. Precision of different algorithms used to recognize RToL traffic

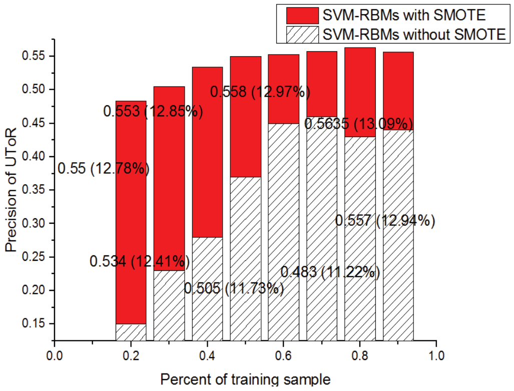

Figure 7. Comparison of precision when classified UToR traffic between method with SMOTE and without SMOTE

The new deep learning method discussed above has several benefits over traditional techniques used in deep learning. To start with, it provides accurate information on anomalous behavior since it can pinpoint the main problem or better still the origin of the intrusion. Furthermore, the system relies on data patterns to detect problems, which would otherwise be difficult to notice through use of human analysts. Additionally, the approach is able to evaluate big sets of data, which is time consuming in traditional deep learning methods. The new technique enables organizations to develop better strategies on network security since the new techniques are more certain than traditional approaches to machine learning. The major advantage to deep coding is its ability to adapt to changing contexts concerning data that ensures the technique conducts exhaustive data analysis.

VII. Conclusion

The use of deep learning has in recent times gained prominence due to its effectiveness in evaluating network security. Of importance, the system has enabled the exhaustive and conclusive assessment of network security. Notably, traditional methods of network security are increasingly failing to function effectively due to increased processing of data. Nonetheless, deep learning has revolutionized the evaluation of challenges in network security. The system uses several approaches to detect abnormalities in the system that include anomaly detection, traffic identification. Nonetheless, the system faces certain limitations that include sanctity of data used to generate inputs and outputs. Similarly, new methods of deep learning are gaining traction due demand for faster and efficient data assessment. Deep belief and deep coding techniques have enabled the analysis of large data sets and deeper system analysis respectively.

References

[1] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, vol. 18, July 2006, pp. 1527-1554, doi:10.1162/neco.2006.18.7.1527.

[2] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur, “Bayesian event classification for intrusion detection,” Proc. 19th Annual Computer Security Applications Conference, Dec. 2003, pp. 14-23, doi:10.1109/CSAC.2003.1254306.

[3] C. Sinclair, L. Pierce, and S. Matzner, “An application of machine learning to network intrusion detection,” Proc. 15th Annual Computer Security Applications Conference, Dec. 1999, pp. 371-377, doi:10.1109/CSAC.1999.816048.

[4] J. Zhang and M. Zulkernine, “A hybrid network intrusion detection technique using random forests,” Proc. First International Conference on Availability, Reliability and Security (ARES'06), April 2006, pp. 8-16, doi:10.1109/ARES.2006.7.

[5] J. Yang, J. Deng, S. Li, and Y. Hao, “Improved traffic detection with support vector machine based on restricted Boltzmann machine,” Soft Computing, vol. 19, Dec. 2015, pp. 1-12, doi:10.1007/s00500-015-1994-9.

[6] (October 2007). KDD Cup 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[7] S. H. Adil, *et al.*, “An improved intrusion detection approach using synthetic minority over-sampling technique and deep belief network,” Frontiers in Artificial Intelligence and Applications, vol. 265, pp. 94-102, doi:10.3233/978-1-61499-434-3-94.

[8] J. R. Quinlan, “C4. 5: Programs for machine learning,” Machine Learning, vol. 16, Sep. 1993, pp. 235-240, doi:10.1007/BF00993309.

[9] G. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” Proc. UAI'95 Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338-345, http://dl.acm.org/citation.cfm?id=2074158.2074196.

[10] C. Chang and C. Lin, “LIBSVM: A library for support vector machines,” 2001, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm