IJCSN International Journal of Computer Science and Network, Volume 5, Issue 1, February 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Comparative Study of Data Mining and Machine Learning Approach for Anomaly Detection

[1] **Sunil M. Sangve** , [2] **Ravindra C. Thool**

[1] Computer Department, Savitribai Phule Pune University , ZCOER
Pune, Maharashtra , India

[2] Computer Science and Engineering Department, SRTMU,SGGSIE &T,
Nanded, Maharashtra, India

Abstract - The intrusion detection systems (IDSs) have attracted more researchers from last two decades. The much more work has been done in IDS. But still, there are some problems remain unsolved like false positive rate and detection accuracy. The various approaches are used in developing IDS; some of these are data mining, machine learning, statistic-based, and rule-based approaches. In this paper, we compare the data mining and machine learning approach for detection of anomaly. We have also discussed the challenges in the intrusion detection system. In studied approaches, some papers used both data mining and machine learning approach for developing system, called as hybrid approach.

Keywords - *Intrusion Detection Systems (IDSs), Data Mining, Machine Learning, Challenges.*

1. Introduction

The anomaly detection is a problem of identifying data points and patterns from given dataset which is also known as outlier detection. The anomaly detection technique applied in various domains such as credit card fraud detection, financial turbulence detection, virus or system intrusion discovery, and network monitoring. The anomaly detection is a binary classification problem with two labels normal and anomaly [11]. The intrusion detection system uses the two stages in i. e training and testing.

There are three techniques of anomaly detection depending upon labels used. The supervised anomaly detection uses the classification with labeled instances (assign normal or anomaly). The semi-supervised anomaly detection technique develop the model which represents normal behavior from given dataset and then test the instances to be generated by learnt model. The un-supervised anomaly technique is used to detect anomalies in an unlabeled test dataset and make the assumption that majority of instances in dataset are normal [8]. The anomaly is defined as the patterns of data that do not give normal behavior. There are various types of anomaly [5].

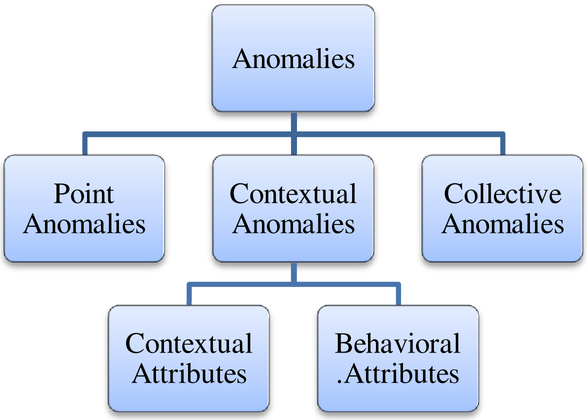


Fig. 1 Types of Anomaly

1.1 Point Anomalies

The data instance considered as anomalous with respect to remaining data. The real time example is credit card fraud detection. Consider dataset of individual's credit card transaction containing one feature (amount spent). If the amount spent transaction is more compared to normal range of expenditure for that person, then it will be a point anomaly.

1.2 Contextual Anomalies

If the data occurrence is abnormal for specific context and not for other, known as contextual anomalies that determines the position of an instance on the entire sequence. It has two types' contextual attributes and behavioral attributes. The contextual attributes are used to determine the context for that instance, for example time-series data, time is a contextual, and the behavioral attributes specify the non-contextual features of an instance. The example is a spatial data set which describes the average rainfall of the world, the total rainfall at any place is a behavioral attribute.

1.3 Collective Anomalies

If a collection of related data instances is anomalous with respect to the entire data set is known as Collective anomalies. For example, KDD Cup 99 dataset consists of a vector with 41 attributes like source packets, destination packets, protocol, f lag etc.

2. Our Contribution

We have studied the two approaches for anomaly detection i.e. data mining and machine learning techniques along with benefits and limitations. We have studied the literature that uses the hybrid approach. The data mining approach is used for large dataset. If large dataset divides into small training dataset, it is possible to decrease the time and computational complexity. The data mining anomaly detection gives the supervised, semi-supervised and unsupervised anomaly detection methods. The machine learning is used to train the system without human interference. If both the approaches are used together then system will become more efficient and effective.

3. Data Mining Approaches in IDS

There are various approaches used in intrusion detection system like statistic-based, data mining, machine learning, and heuristic-based approach. Amongst these approaches, we are discussing data mining and machine learning approaches.

The use of data-mining in IDS:

1. The dataset is very large and contain valuable patterns which later discovered automatically.
2. Manual updating of data is not easy. Thus, classification and clustering algorithms are used in training stages.

Jungsuk Song et al. [1], proposed a more practical unsupervised anomaly detection system. The unsupervised anomaly detection technique construct the IDS models with unlabeled training data automatically and detects the unknown attacks for example, 0-day attacks. They resolved the problem of tuning and optimization of parameters which are required for building of IDS process. The results are calculated on real traffic data, received from Kyoto university honeypots.

3.1 The Adaptive Detection Approach

The adaptive detection approach was proposed by Ji zang et al. [2]. The importance of real-life big data sets have increased and gained a lot of research interest in data mining. From large dataset, it is very difficult to extract required knowledge and patterns for detection of network traffic anomalies. The authors studied this problem and proposed new method based on outlier detection, called as adaptive stream projected outlier detectors (A-spot). They used 1999 KDD Cup dataset and consider the parameter for training data generation, anomaly classification and false positive reduction. The A-spot is categorized in two steps: learning and detection.

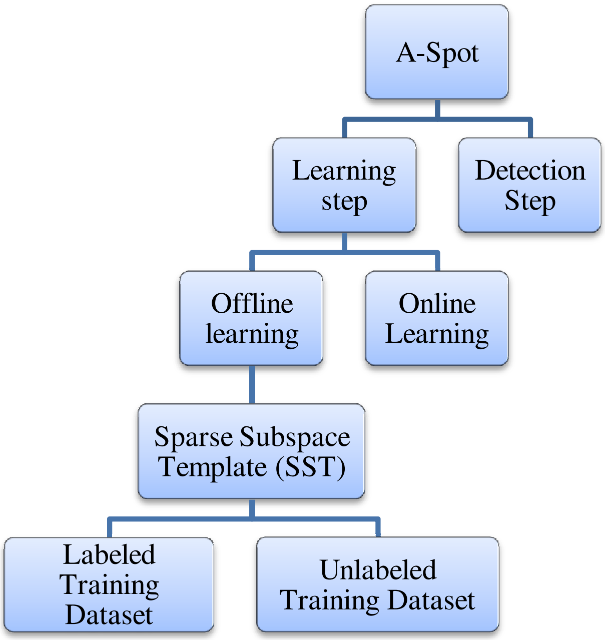
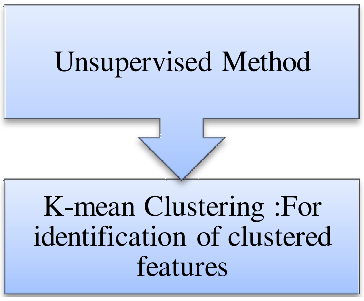
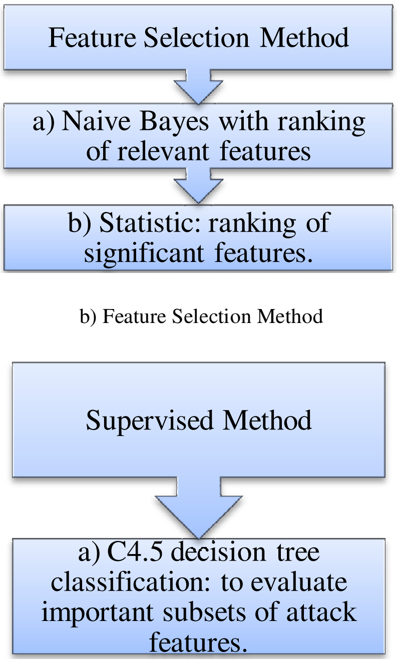


Fig. 2 Adaptive Detection Approach



a) Unsupervised Method


Feature Selection Method
a) Naive Bayes with ranking of relevant features
b) Statistic: ranking of significant features.
Supervised Method
a) C4.5 decision tree classification: to evaluate important subsets of attack features.

b) Feature Selection Method

c) Supervised Method

Fig. 3 Effect-based feature identification method (combing the above methods given in Fig 3 a),b) and c))

A-spot uses Sparse Subspace Template (SST), created by using both labeled and unlabeled training data. As SST developed, A-spot start to identify the anomalies which are arrived from data. The incoming data update the data synopsis i. e data characteristics of anomaly detection. The data is called anomaly if data synopsis is less than predefined threshold. The outlier repository used to store detected anomalies.

3.2 Effect-based Feature Identification

Proposed by Panos et al. [3], in which anomaly detection technique (ADT) with K-mean, Naïve-Bayes feature selection and C4.5 decision tree for increasing detection accuracy and creates awareness between cyber network operators. This approach is implemented using three stages:

1. Unsupervised method
2. Feature selection method
3. Supervised method

The unsupervised method uses K-mean clustering for identification of clustered features. The feature selection method uses Naïve Bayes with ranking of relevant features and statistic for ranking of significant features.

3.3 Hybrid K-mean Algorithm in Content-centric Networks

Amin Karami et al. [4] proposed a novel fuzzy approach for anomaly detection. The K-mean algorithm has some limitations in selection of cluster centroid. The Fuzzy approach uses the following two steps for anomaly detection:

1. To determine optimal number of clusters, hybridization of particle swarm optimization (PSO) and K-mean clustering together, in training phase.
2. In second phase, fuzzy approach is used by taking two distance–based methods, as classification and outlier, in detection phase. If optimal result is greater than threshold then data is anomalous otherwise normal. They calculated detection rate and false positive rate.

The main challenges in developing IDS are false positive rate, false negative rate and data overload. The development of unique signature is very hard task. If instances are normal and detected as abnormal then they are called as false positive rate. If system is not generated the alarm when attack takes place, known as false negative. The data overload does not relate to detection, but it is very important that analyst can consider all types of data very effectively.

The problems in IDS using data mining approaches are [5]:

1. To define a normal region which consist of all possible normal behaviors.
2. The boundary between normal and abnormal also not precisely defined. So, anomalous observation which is close to boundary may be normal.
3. The labeled data which is available for training and testing in anomaly detection is also a big issue.
4. The data contains a noise and looks similar to actual anomalies. It is very difficult to distinguish and remove.

In the research of IDS, the various factors are considered like nature of data, availability of labeled data, types of anomalies detected.

4. Machine Learning Approaches

4.1 Advantages of Machine Learning Approach

1. The signature-based IDS depend on humans to create, test and deploy signature manually.

2. To generate new signature may require hours of time or days. To provide human independent solution machine learning approach is used.
3. Machine learning approach has the ability to develop the system that learns from previous data and give the feedback for unknown data.
4. It consists of neural network, support vector machine, genetic algorithm, fuzzy logic, Bayesian network, decision tree.
5. The machine learning approach popular for many real time problems, based on both explicit and implicit model.

4.2 Neuro-Fuzzy Intrusion Detection System

Fuzzy set theory is used for approximate reasoning instead of prediction. The features are considered as fuzzy variables. It is very useful in probes related attacks but it requires high resource consumption. The fuzzy logic specified as if {condition} then consequence, where condition= fuzzy variable, consequence= fuzzy set.

Varun Chandola et al. [6], proposed hybrid approach for anomaly detection. They combined two techniques i. e Artificial Neural Network and Fuzzy inference system and used SNORT (Libcap- based sniffer and logger) tool to perform real time traffic analysis with DARPA 1998 dataset.

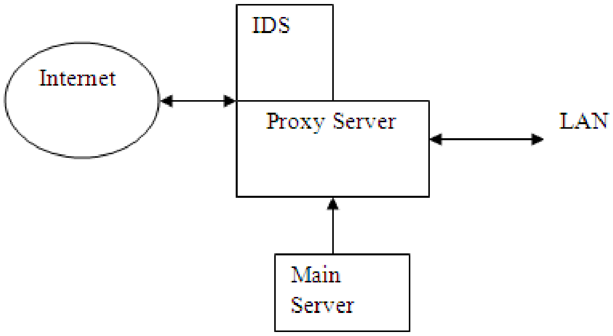


Fig. 4: Neuro-Fuzzy IDS [6]

The approach consists of proxy server, internet, LAN, IDS components. The LAN gives the connection between nodes with Ethernet topology; internet is a gateway to the external world. The IDS learns the new types of attacks without disturbing previously trained knowledge. IDS update the technology by using if-then fuzzy rules. The proxy server used to identify application or a packet and if that particular service available then allowed to pass through. Because of this, there is no direct connection between the un-trusted and trusted system. It performs address translation, mapping of IP address to valid IP address.

4.3 Self Organizing Maps (SOMs) and Particle Swarm Optimization (PSO)

M. Lofti Shaherza et al. [7], proposed an unsupervised neural network approaches by combing self organizing maps (SOMs) and particle swarm optimization (PSO).

In the SOM, the single layers of neurons are arranged in rectangular array. It checks the input patterns and calculates weight to match input. The neurons which has more similarity, is declared as winner. The SOMs reduced the information with keeping most important topological relationships and trained with unsupervised learning. It is very efficient in handling large dataset.

The PSO are developed through simplified social model simulation. The particle is initializing with random position and velocities. It is effective in non-linear optimization problems and also easy to implement. The PSO applied in various domains like genetic algorithm, fuzzy system control and ANN training.

Genetic Algorithm includes evolutionary algorithm techniques. It is capable for deriving classification rules and selects optimal parameters for IDS.
Steps:

1. Collect the information from specific network.
2. IDS apply trained GA along with classification rules.
3. Use the set of rules to classify normal and abnormal patterns.

4.4 Entropy and Support Vector Machine Approach

Basant Agarwalet et al. [10], hybrid approach for anomaly detection using support vector machine (SVM) and entropy of network features has been proposed. The normalized entropy values of different network features are measured. The SVM model is trained for identification of normal and anomalous traffic. For detection MIT Lincoln Laboratory dataset is used (DARPA, 1999).

4.4.1 Entropy based Intrusion Detection System

Entropy is a measurement of randomness. If entropy of network features deviate from a threshold value, then it indicates abnormality in the network traffic i.e. anomaly in the network traffic. The network features considered for detections are packet size, source IP address, source port number, destination IP address, destination port number, and packet type (ICMP, TCP, and UDP).

4.4.2 Support Vector Machine Based Intrusion Detection

It is a useful classification technique and gives prediction of normal or anomaly traffic. The network features considered for detections are source IP address, source port number, destination IP address, destination port number, , packet size, different packets with same size, packet type (ICMP, TCP, and UDP). By combining these two techniques, system gives better result for detection of anomaly and false alarm rate.

Ujwala Ravale et al. [9], proposed hybrid technique combines data mining approaches like K Means clustering algorithm and RBF kernel function of Support Vector Machine as a classification module.

This technique decreases the number of features associates with data instances. They calculated detection accuracy using KDD Cup 99 dataset. In first stage, they group the data instances depending on their behaviors by utilizing K-Means clustering and in next stage, RBF kernel is used for classification of anomaly and normal data packets. The K-Means clustering technique reduces large heterogeneous dataset to a number of small homogeneous subsets.

Table 1: Machine Learning Methodologies.

References	Processing Strategy	Detection methodology	Dataset used	Network traffic	Detection Nature
S. C. Lee and D. V. Heinbuch [12]	Centralized	Neural network based intrusion detection.	Simulated data	Packet based	Non-real time
J. Q. Xian et al. [13]	Centralized	Intrusion detection based on clonal selection clustering algorithm	KDD Cup 99	Packet based	Non-real time
M. Amini et al. [14]	Centralized	Unsupervised neural network intrusion detection	KDD Cup 99, real life	Packet based	Real time
W. Chimphlee et al. [15]	Centralized	Detection using Fuzzy Rough Clustering	KDD Cup 99	Packet based	Non-real time
Liu et al. [16]	Centralized	Neural network based intrusion detection.	KDD Cup 99	Packet based	Non-real time
R. C.Chen et al. [17]	Centralized	Rough set and support vector machine is used for intrusion detection	DARPA 98	Packet based	Non-real time
S. Mabu et al. [18]	Centralized	Fuzzy Class-Association-Rule Mining Using Genetic Network Programming	KDD Cup 99	Packet based	Non-real time
A. Visconti and H. Tahayori [19]	-	type-2 fuzzy set	Real life	Packet based	Non-real time
F. Geramiraz et al. [20]	-	Fuzzy controller	KDD Cup 99	Packet based	Non-real time

5. Conclusions

Considering the studied literature, it is clear that in order to have the capacity to secure a system against the novel attacks, the anomaly based intrusion detection is the best way. We have discussed the data mining and machine learning approach for anomaly detection. The data mining approaches are used for clustering and classification to divide the large dataset so that processing and computational complexity will reduce. The machine learning approach trains the system and gives the prediction in testing stage. The machine learning has many advantages in anomaly detection without human interference. Depending on labels used in input dataset, the anomaly detection is classified as supervised, semi-supervised, and un-supervised. We are also focusing on combining the best features from data mining and machine learning approach and will propose the hybrid approach which gives better result than discussing methodology.

References

[1] Jungsuk Song, Hiroki Takakura, Yasuo Okabe, Koji Nakao, “Toward a more practical unsupervised anomaly detection system”, Information Sciences 231 (2013) 4–14

[2] Ji Zhang, Hongzhou Li, Qigang Gao, Hai Wang, Yonglong Luo, “Detecting anomalies from big network traffic data using an adaptive detection approach ”, Information Sciences 318 (2015) 91–110

[3] Panos Louvieris, Natalie Clewley, Xiaohui Liu, “Effects-based Feature Identification for Network Intrusion Detection”, Neurocomputing121(2013)265–273.

[4] Amin Karami, Manel Guerrero-Zapata, “A Fuzzy Anomaly Detection System based on Hybrid PSO-Kmeans Algorithm in Content-Centric Networks”, Neurocomputing149 (2015)1253–1269.

[5] Sampada Chavan, Khusbu Shah, Neha Dave, Sanghamitra Mukherjee, Ajith Abraham and Sugata Sanyal, “Anomaly Detection: A Survey”, ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI = 10.1145/1541880.1541882.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar, “Adaptive Neuro-Fuzzy Intrusion Detection Systems”, IEEE(ITCC 2004), Proceedings of ITCC 2004, Vol. 1, April, 2004, Las Vegas, Nevada, USA pp. 70-74.

[7] M. Lotfi Shahreza, D. Moazzami, B. Moshiri, M.R. Delav,” Anomaly detection Using a Self-Organizing Map and Particle Swarm Optimization”, Scientia Iranica D (2011) 18 (6), 1460–1468.

[8] Xiaojin, Zhu ‘‘Semi-supervised learning literature survey’’, Computer Sciences TR 1530, University of Wisconsin–Madison, Last modified on July 19 (2008).

[9] Ujwala Ravale, Nilesh Marathe, Puja Padiya, “Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function”, Procedia Computer Science 45 (2015) 428 – 435

[10] Basant Agarwal, Namita Mittal, “Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques” ,Procedia Technology 6 (2012) 996 – 1003.

[11] Zhe Yao, Philip Mark, and Michael Rabbat, “Anomaly Detection Using Proximity Graph and PageRank Algorithm” ,IEEE transactions on information forensics and security, vol. 7, no. 4, august 2012.

[12] S. C. Lee and D. V. Heinbuch, “Training a Neural-network based Intrusion Detector to Recognize Novel Attacks”, IEEE Trans. Syst. Man Cybern. A, vol. 31, no. 4, pp. 294–299, 2001.

[13] J. Q. Xian, F. H. Lang, and X. L. Tang, “A Novel Intrusion Detection Method based on Clonal Selection Clustering Algorithm”, in Proc.(ICMLC) .USA: IEEE Press, 2005, vol.6.

[14] M. Amini, R. Jalili, and H. R. Shahriari, “RT-UNNID: A Practical Solution to Real-Time Network-based Intrusion Detection using Unsupervised Neural Networks”, Computers & Security, vol. 25, no. 6, pp. 459–468,2006.

[15] W. Chimphlee, A. H. Abdullah, M. S. M. Noor, S. Srinoy, and S. Chimphlee, “Anomaly-Based Intrusion Detection using Fuzzy Rough Clustering”, in Proc (ICHIT), vol. 01. Washington, DC, USA: IEEE Computer Society, 2006, pp. 329–334.

[16] G. Liu, Z. Yi, and S. Yang, “A Hierarchical Intrusion Detection Model based on the PCA Neural Networks”, Neurocomputing, vol. 70, no. 7-9, pp. 1561–1568, 2007.

[17] R. C. Chen, K. F. Cheng, Y. H. Chen, and C. F. Hsieh, “Using Rough Set and Support Vector Machine for Network Intrusion Detection System”, in Proc. (FACIIDS). Washington, DC, USA: IEEE Computer Society, 2009, pp. 465–470.

[18] S. Mabu, C. Chen, N. Lu, K. Shimada, and K. Hirasawa, “An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming”, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 41, no. 1, pp. 130–139, 2011.

[19] A. Visconti and H. Tahayori, “Artificial Immune System Based on Interval Type-2 Fuzzy Set Paradigm”, Applied Soft Computing, vol. 11, no. 6, pp. 4055–4063, September 2011.

[20] F. Geramiraz, A. S. Memaripour, and M. Abbaspour, “Adaptive Anomaly-Based Intrusion Detection System Using Fuzzy Controller”, International Journal of Network Security, vol. 14, no. 6, pp. 352–361, 2012.