Enhancing Cloud of Things performance through Intrusion Detection via machine learning

Sami MAHFOUDHI

Department of Management Information Systems, Qassim University

Abstract
Recently, Internet of things (IoT) has become one of the hot topics of research. Things are, in most cases, deployed in unmonitored fields. So, ensuring the reliability of collected data is becoming a challenging issue. One of the most common problem affecting IoT is caused by intrusion, which could alter, delete or modify data collected by things. It could affect the whole functionality of IoT by causing faulty situations like taking wrong decisions. On one hand, the intrusion is among the hardest attack that could to be detected. On the other, Artificial Intelligence (AI) tools are powerful and emerging techniques that could be used to achieve this purpose. In this paper, we propose to use classification techniques to deal with the problem of intrusion detection. More precisely, we applied a set of classification tools on a real IoT dataset to detect intrusion. The comparison of classification results is shown through an experimental study.

Keywords
IoT, Security, Intrusion detection, machine learning, classification techniques.

1. Introduction

Recently, the number of internet users all over the world is becoming bigger than the human population in the planet. This is due to the great number of intelligent devices connected to the internet. These autonomous devices are called things, and connecting them to the internet leads to the appearance of the term Internet of Things IoT [1].
IoT has attracted the attention of community as a promoting field. Indeed, IoT are used in several domains including daily life, medicine, traffic mentoring, forest controlling, etc. This is due to the low cost of things because there is no need to build new networks for them since they are already using existent internet connection.
The use of IoT is facing many challenges making it a hot topic of research [2]. Some of these challenges are data management, data mining, and specially the security issue for which we are trying to answer questions like: how to protect collected data and how to guarantee its originality regarding confidentiality, integrity and availability.
Things are, in most cases, deployed in unmonitored fields. This makes them vulnerable to many attacks and threats. In addition to basics threats such us software or hardware malfunctioning, things are prone to many attacks like man-in-the-middle attacks, denial of services, black hole attacks, green hole attacks, etc. Attacks aiming to modify or alter collected data are difficult to discover. Consequently, designing or using Intrusion Detection System is high priority to ensure good functionality of IoT [3].
Many techniques and methods have been proposed by research community to deal with intrusion detection in IoT [3]. The aim of this paper is to deal with this problem using machine learning techniques employing a set of classification tools. Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), as powerful tools of classification, are used and compared in this paper to detect intrusion in IoT. The comparison is performed on a real data set, and shown through a detailed experimental study. This study is based on rate of detection, rate of positive alert and rate of negative alert.
The following is organized as : section 2 outlines the most important and recent works related to ID in IoT. Section 3 makes an overview on SVM classifier. In section 4, our contribution is detailed. The experimental study is discussed in section 5. Finally, section 6 is a conclusion of this paper.

2. Related work

In this section, recent and important techniques proposed by research community dealing with detection intrusion in IoT are outlined. According to [3], intrusion detection methods in IoT could be classified into four types which are: (1) signature-based, (2) anomaly-based, (3) specification-based, and (4) hybrid methods.
The signature-based ID techniques are stored in database among other attack signatures. The IDS triggers alerts if the system or the network behaviour is similar or seems like a stored one. In [4] the authors proposed a signature based IDS aiming the detection of DoS attacks in networks that use 6LoWPAN as addressing system. To confirm that an attack is happen, the IDS sends alerts to the DoS manger to achieve this task. In [5], the authors adopt an optimization techniques to avoid unnecessary matching between packet and attacks signature. The avowedness is due to IoT resources limitations.
For IDS, other approaches based on anomaly is issued in other researches. In [6], authors used this kind of IDS, the proposed technique is based on monitoring first node

Manuscript received May 5, 2019
Manuscript revised May 20, 2019

neighboring characteristics like packet size and data rate. This helps segregate normal behavior and abnormal one. But, the authors did not take into consideration the resource limitation of IoT devices. Another research work which can be considered as animal-based IDS is proposed in [7]. The proposed IDS is designed to discover wormhole attacks targeting IoT devices.

The third kind of IDS is using specification based approach. Authors of [8] presented a technique for IDS based on specification. They concentrate on detecting attacks on RPL based network topology. Their contribution is based on a finite state machine aiming to manage network in order to distinguish malicious behavior. In [9], the presented techniques are based on rules creation by network administrator to detect future attacks. The IDS alerts the Event Management Systems (EMS) in case one of these rules is violated.

Last kind of IDS that can be outlined here concerns the hybrid based approach. The aim of this kind of IDS is to take into consideration IoT devices as resources constrained devices in term of computing capacities and storages. Many research works combine signature-based approach and anomaly-based techniques. In [10], the author propose a new framework to evaluate their contribution which mixes signature and anomaly based IDS. The authors show that the combination of the two approaches outperforms the fact of using each approach separately in term of discovering attacks. In [11], the authors use also the same hybrid concept but this time to deal with sinkhole attack based on a probability to make decisions on node trust.

As a summary of this related works section, we can notice that, although the machine learning techniques specially classification method are powerful tools, but they were not widely used to deal with IDS in IoT. The aim of this paper is to compare the use of DT, RF and SVM classifier to achieve this goal. Next, an overview of this set of tools is performed.

3. Classification Techniques: An overview

Among multiples techniques of classification, SVM, DT and RF are chosen in our paper because they are the most important in literature. In addition, they are widely used in classification problem while they provide significant results.

3.1. SVM classifier

SVM classifier, as a machine learning tool [12], has been widely used to solve problems that need classification. In this section, an overview of SVM is outlined.

The aim of classification techniques is to separate data into diverse classes [13]. First, SVM uses a set of training data in order to pre-label classes. Next, for an inputting dataset, this technique can predict their class membership based on the previously performed training.

The separations between classes in SVM are called hyperplanes. The identification of hyperplanes is performed according to support vectors. Figure 1 shows two classes separated by a hyperplane defined according to support vectors. SVM classifier can be used for both types of data: linear and nonlinear.

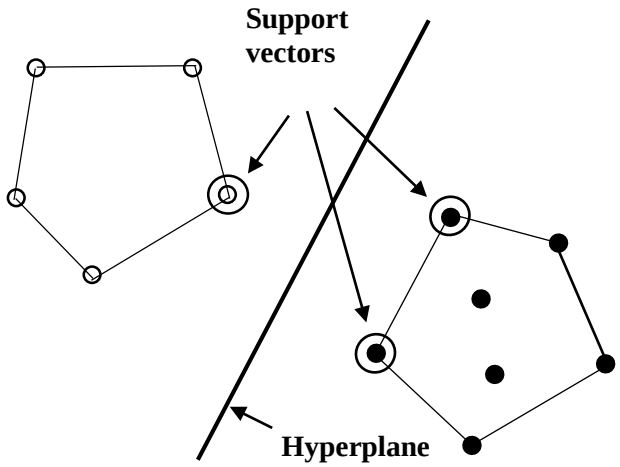

Fig. 1 Support vectors and hyperplane [13]

In case the data can be separated linearly the equation of hyperplane can be define as following:

$$WX + b = 0$$

Where:

W=(w0,w1,…) is representing the vector of weight*;* *X= (x1,x2,…)* is the training data and *b* is a scalar bias.

According to Lagrangian formulation and using Karush-khun-Tucker condition, the previous equation can be reformulated as bellow, to describe the maximum margin hyperplane as below:

$$x = \sum_{i=1}^{n} \alpha_i \, y_i a(i).\, a + b$$

Where n is the number of support vectors, b and α_i are learned parameters, a(i) represents the vector of SV, and i instantiates a support vector.

3.2. DT classifier

Decision Tree classifier [14] consists in dividing the space in a recursive way. It is a direct tree containing a root node in addition to internal node and leaf node. There is no

entering node for root and no leaving node from leaf. Tests on attribute are represented in internal node. The tests result of test is represented by the parting edge. Furthermore, the instance space is partitioned to subspaces in internal node resulting from applying discrete function on the input values. So that, DT divides recursively the training set until the sample parts are totally or mostly compromised sample parts from one class. The procedure of partition is continued until obtaining partition which is small in size or perfect. Figure 2 is an example of a DT.

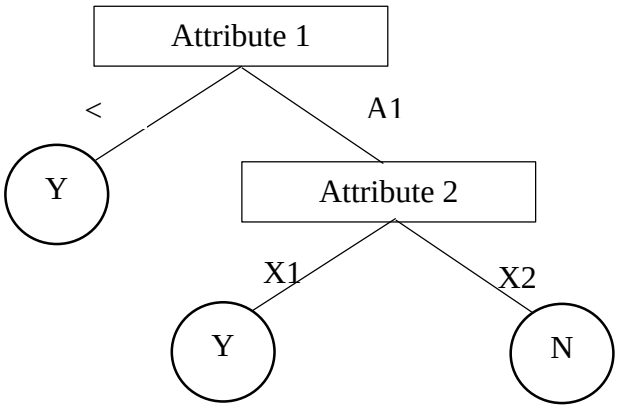

Fig. 2 Decision Tree

3.3. RF classifier

Random Forest classifier [15] is a machine learning tool used for classification as well as for regression. It is based on constructing multiple decision trees at the training phase. Next, the class is selected. The aim of RF is to improve DT.

4. Contribution

In this section, the contribution of this paper is outlined. As previously mentioned, the aim of this paper is to apply a set of classifiers on a real dataset [16,17] for intrusion detection. The measurements of our technique performance are performed according to metrics defined in the next subsection, and analyzed in the experimental study subsection.

4.1. Accuracy

To measure the performance of an intrusion detection techniques many metrics can be used to describe their success. In this paper we compare the three classifiers tools depending on their accuracy. The accuracy is the percentage of testing sets correctly classified.

4.2. Experimental study

The data set includes many types of attacks which are summarized and presented in table1.

Table 1: Attack type and example.

Attack type	Example
Denial of service (DoS)	Apache2, Smurf,Neptune, Back,…
Remote t local (R2L)	Guess_Password, FTP_write, Named, ….
User to root (U2R)	HTTPtunneel, Rootkit, Xtem,Ps,…
Probe	Saint, Satan, Mscan

The data set of [16] is prepared according to the following steps:

- Step 1: Collecting Data: getting a data frame from the KDD train and test and column labels csv files.
- Step 2: Merging Data: merging all data frames to one data frame (df). Then, looking for Nan values in the different columns. Finally, drop duplicates if any from (df).
- Step 3: Normalizing Data: scaling numeric data with a min-max scalar (values become between 0 and 1). Next, applying hot encoding for categorical columns (more than two categories). Then, adding target column to the end of the data frame. Finally, changing target column type to categorical.

After the three previous steps, plotting the cumulative summation of the explained variance to select the desired number of components in figure3.

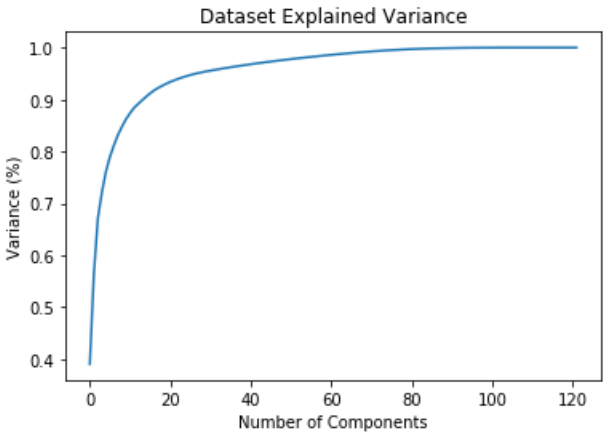

Fig. 3 Dataset Explained Variance.

- Step4: Reduce dimension of selected features using PCA. The number of features (components) becomes 10, and the total variance from PCA components is 0.8621156371639466.

After step4, the dataset is ready and we can apply the three classifier methods. The accuracies of SVM classifier, DT classifier, and RF classifier are given in figure 4 and table1.

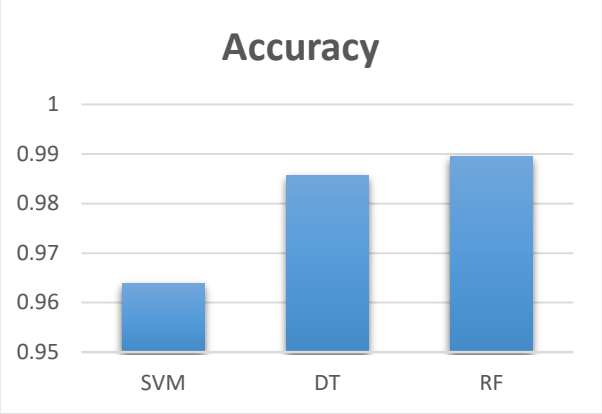

Fig. 4 Accuracy comparison

According to figure 4 and table 2, the best accuracy is one of RF classifier which is better than the accuracy of SVM classifier by an average of 2.65%, and lightly exceeds the accuracy of DT classifier by an average of 0.037%.

Table 2: Accuracy values.

SVM classifier	DT classifier	RF classifier
96.37%	98.56%	98.93%

According to table2, RF classifier outperforms DT classifier. Indeed, the first technique is based on constructing multiple decision trees at the training phase. So, in most case RF classifier performs better than DT classifier which is confirmed in this case. While accuracy of SVM is less than DT and RF due to the nature of dataset itself. In most cases, SVM classifier performs worse when standardizing data.

5. Conclusion

In this paper the intrusion detection problem in IoT is studied. Due to IoT devices characteristics, as constrained resources, many existing techniques with high complexity in term of storage and time cannot be applied. Therefore, applying lightweight methods is highly recommanded in this context. Consequently, in our contribution a set of classifier techniques are used and compared as a powerful learning tools to deal with IoT IDS. According to the performed experimental study with a real dataset, we have shown our proposition effectiveness in term of intrusion detection. This comparison study shows that RF classifier outperforms both SVM and DT classifiers.

Acknowledgment

The authors thank and acknowledge the scientific research deanship at Qassim University for their financial support during the academic year 2017/2018 under research grant reference number 5234-CBE-2018-1-14-S.

References

[1] Ray, Partha Pratim. "A survey on Internet of Things architectures." Journal of King Saud University-Computer and Information Sciences 30.3 (2018): 291-319.

[2] Lee, In, and Kyoochun Lee. "The Internet of Things (IoT): Applications, investments, and challenges for enterprises." Business Horizons 58.4 (2015): 431-440.

[3] Zarpelão, B. B., Miani, R. S., Kawakani, C. T., & de Alvarenga, S. C. (2017). A survey of intrusion detection in Internet of Things. Journal of Network and Computer Applications, 84, 25-37.

[4] Kasinathan, P., Costamagna, G., Khaleel, H., Pastrone, C., & Spirito, M. A. (2013, November). An IDS framework for internet of things empowered by 6LoWPAN. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (pp. 1337-1340). ACM.

[5] Oh, D., Kim, D., & Ro, W. W. (2014). A malicious pattern detection engine for embedded security systems in the Internet of Things. Sensors, 14(12), 24188-24211.

[6] Thanigaivelan, N. K., Nigussie, E., Kanth, R. K., Virtanen, S., & Isoaho, J. (2016, January). Distributed internal anomaly detection system for Internet-of-Things. In Consumer Communications & Networking Conference (CCNC), 2016 13th IEEE Annual (pp. 319-320). IEEE.

[7] Pongle, Pavan, and Gurunath Chavan. "Real time intrusion and wormhole attack detection in internet of things." International Journal of Computer Applications 121, no. 9 (2015).

[8] Le, Anhtuan, Jonathan Loo, Kok Keong Chai, and Mahdi Aiash. "A specification-based IDS for detecting attacks on RPL-based network topology." Information 7, no. 2 (2016): 25.

[9] Amaral, J. P., Oliveira, L. M., Rodrigues, J. J., Han, G., & Shu, L. (2014, June). Policy and network-based intrusion detection system for IPv6-enabled wireless sensor networks. In Communications (ICC), 2014 IEEE International Conference on (pp. 1796-1801). IEEE.

[10] Krimmling, J., & Peter, S. (2014, October). Integration and evaluation of intrusion detection for CoAP in smart city applications. In Communications and Network Security (CNS), 2014 IEEE Conference on (pp. 73-78). IEEE.

[11] Cervantes, C., Poplade, D., Nogueira, M., & Santos, A. (2015, June). Detection of sinkhole attacks for supporting secure routing on 6LoWPAN for Internet of Things. In IM (pp. 606-611).

[12] Zidi, Salah, Tarek Moulahi, and Bechir Alaya. "Fault detection in wireless sensor networks through SVM classifier." IEEE Sensors Journal 18.1 (2017): 340-347.

[13] Support vector machines: concise technical overview, https://www.kdnuggets.com/2016/09/support-vector-machines-concise-technical-overview.html

[14] Gupta, Gaurav. "A self-explanatory review of decision tree classifiers." Recent Advances and Innovations in Engineering (ICRAIE), 2014. IEEE, 2014.
[15] Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493-507.
[16] Data Set, https://www.unb.ca/cic/datasets/nsl.html
[17] Botes, F. H., Leenen, L., & De La Harpe, R. (2017, June). Ant colony induced decision trees for intrusion detection. In ECCWS 2017 16th European Conference on Cyber Warfare and Security (p. 53). Academic Conferences and publishing limited.