Method to implement K-NN machine learningto classify data privacy in IoT environment

QahtanMakki Shallal[1], ZaidAlaa Hussien[2], Alaa Ahmed Abbood[3]
[1,2]Management Technical College of Basra, Southern Technical University, Iraq
[3]Faculty of Business Informatics, University of Information Technology and Communications, Iraq

Article Info

Article history:

Received Feb 12, 2020
Revised Apr 10, 2020
Accepted Apr 24, 2020

Keywords:

Asymmetric RSA algorithm
Confidentiality
Internet of things
Learning machine
Symmetric AES algorithm

ABSTRACT

Internet of Things technology allows many devices to connect with each other. The interaction could be between humans and devices or between devices itself. In fact, the data are traveling between the devices through the media within the boundary, and it could be traveling outside the boundary when it required to be analyzed or stored in the cloud through the internet. Due the transmission media and internet, the data are vulnerable to attacks. Thus, the data need to be encrypted strongly for the purpose of protection. Usually, most of the encryption techniques will consume computer resources. In this work, we divide the data that are used in the IoT environment into three levels of sensitivity which are low, medium and high sensitive data to leverage the computer resources such as time of encryption and decryption, battery usage and so on. A framework is proposed in this work to encrypt the data depends on the level of sensitivity using the machine learning K nearest neighbors (K-NN).

Corresponding Author:

QahtanMakkiShallal,
Department of Information Technology,
Southern Technical University,
Iraq.
Email: qahtan.makii@stu.edu.iq

1. INTRODUCTION

With the growth of using IoT technology and its reputation, the users and devices of IoT are increased gradually, also the data traffic becomes vast between devices and users. In fact, The International Telecommunication Union (ITU) informed that the challenges of users' privacy is too necessary for the IoT. The ability to obtain the personal information has noticeably improved by the sensors. Further, to support integrated services, the sensors have to be integrated into vehicles, buildings, and common environments, passed by humans and attached to the animals to be communicated between them on the same site or in a different site. IoT is having many different devices such as GPS, cameras and RFID, the information that belongs to one environment can be collected accurately and integrally, such as the speed of movement, the location of the device and physical signs, like (disease, pulse, blood pressure, etc.) [1-3].

Thus, the IoT ability to gather personal privacy has been grown along with the ability of expansion in IoT technology. IoT has the ability to fetch the interest of hackers for the purpose of political and commercial. Especially as the technologies of IoT are commonly used in a range of industries, military, national defense, and many other interesting areas. Therefore, due to the specific internet virus and hackers, they will harm the environment. The above two points mentioned pose a dangerous threat to the security of personal information in the IoT environment. Unfortunately, protecting privacy is not getting much attention from the researchers. A number of researchers believed that the available technologies of encryption are able to solve the security of privacy issues in the IoT environment [4-6].

Leakage or loss of data will have a negative impact on the trustworthiness of the technology and make people afraid of joining the technology. To ensure the data security of IoT, we must find a way to protect the data during transmission without effecting much the computer resources [7]. When the IoT technology required to send its data to the cloud using an internet connection to be analyzed and stored, a strong encryption technique must be executed on that data. This strong encryption needs to consume time, memory, CPU, and battery. As we know not all data is sensitive to the environment, so dividing the level of sensitivity is important to decrease the usage of resources [8-10].

2. LITERATURE REVIEW

Several privacy protection technologies have been discussed recently, but most are independent and aim to protect specific privacy. They also ignored research on privacy features. IoT applications can include different types of privacy, and their protection is very expensive. Due to the economic cost of implementing the system and the efficiency of implementation, it may not be possible to use all privacy protection techniques in some cases. Therefore, privacy must be classified into different categories so that limited privacy can be protected at a limited cost and technology. In [11], the authors have used hybrid techniques by implementing both RSA and digital signature to protect the data during transmission to the cloud environment. In this work, the digital signature used to ensure the particular message did not alter and has sent by the specific user which is known well to the cloud server. The work in [12] proved that the algorithm of symmetric AES is faster and efficient than the other symmetric algorithms. The author declared that the AES algorithm is appropriate to do encryption/decryption for the data when it resides outside the boundary.

In [13], authors have utilized Diffie key exchange as well as digital signature blended with symmetric AES algorithm to keep the confidentiality of data that is stored in a cloud server. Three schemes of protection used in their work. Initially, Diffie–Hellman algorithm has been used to produce a set of keys for the purpose of key exchange. Moreover, theyusing a digital signature to fulfill the process of authentication. Finally, the symmetric AES has implemented to encrypt and decrypt the data of the user. The proposed method in this work used to do not allow any change in data in the server. The authors in [14], have introduced an approach to encrypting the data which are sensitive by utilizing a two times encryption techniques to ensure the security of data during the transmission period. They used symmetric DES to do encryption on the plaintext and then used asymmetric RSA to encrypt the secret key made by the DES algorithm. Authors had developed a hash-based message authentication code (HMAC) to guarantee the message integrity.

In [15], the authors have been compared many symmetric algorithms which are (AES, DES, 3DES, Blowfish) alongside (RSA, Diffie–Hellman) asymmetric algorithms. Authors detected symmetric algorithms having a high ratio of encryption than asymmetric algorithms. Furthermore, a higher tenability found in asymmetric algorithms compared to symmetric. Additionally, the key length in asymmetric is higher than that exists in symmetric, thus it is too difficult to unlock the codes that used RSA algorithm. Moreover, the mechanism of symmetric algorithms is faster compared to asymmetric. Finally, from the perspective of security, they found that AES is preferable than other symmetric algorithms in the study, and RSA is better than Diffie-Hellman. In [1], authors are investigating the features of privacy information and suggesting two new features of privacy: the universality of privacy and sensitivity of privacy. It suggests ways to classify security levels for privacy information and suggests three security goals for the three privacy levels. Security levels are categorized by 52 privacy items based on big data in queries from "Baidu knows" search engine. In the future, depending on the level of security of IoT privacy, security measures of varying complexity must be implemented to achieve the corresponding security objectives.

In [16] authors reviewed different techniques of security and its challenges from the hardware as well as software aspects to safeguard the data in cloud. Further, it aims to enhance data privacy and security protection for the reliable environment of the cloud. Thus, the authors made a comparative analysis for the existing research papers considering the protection techniques of data privacy and security used in the environment of cloud computing. The authors in [17], proposed a method to encrypt the data depending on their security level. They divide the data into two levels of security which are high or normal sensitive data, the security level of data is determined by machine learning, they used algorithm of K nearest neighbors. The technique of OTP used to ensure user authentication, the authors applied AES-192 algorithm for the normal sensitive. While a hybrid AES-256 and RSA were applied for the high sensitive level. Finally, they used to attached HMAC at the end of the message for the purpose of ensuring the authenticity and integrity of the message. In their research, they measured the encryption/decryption time, memory usage, and throughput.

3. KNN MACHINE LEARNING

The K-nearest neighbor classifier (K-NN) is one of the preparatory supervised classification technique, which every science learner of data should know about it. Fix & Hodges introduced the K-NN classifier algorithm in 1951 for the purpose of accomplishing the task of pattern classification. Knn focuses on the problems of pattern recognition, weather prediction, color reorganization, and many more. The uncomplicated K-NN classifier version is to expect the label of target by finding out the class of the nearest neighbor [17]. The nearest class will easily be identified by calculating the distance like Euclidean distance.

3.1. The KNN algorithm [17, 18]

1) Load the unclassified data.
2) Set K value to your chosen neighbors.
3) For each and every example inside the data.
4) Compute the distance among the current and query example belongs to the data.
5) Add both index and distance of the example into the ordered collection.
6) Organize the ordered group of distances and indices from the smallest to largest value considering the distances.
7) Select the first entries of K from the collection that are sorted.
8) Fetch the labels from the selected entries of K.
9) If regression, get back the K labels.
10) If classification, get back to the method of K labels.

3.2. Use of K-NN algorithm to classify data

K-NN learning machine has been used for many types of research to classification, prediction, estimation, and pattern recognition. It relies on the idea of instance-based learning, which contains a number of data that trained and saved to find the class of unclassified data. Furthermore, K-NN classifier is effective well with the recognition problems [19]. Using the larger K can potentially some similar pixels; as another option, using the smaller K may exclude a range of candidate pixels. In both situations, the classification accuracy will certainly be decreased. In fact, the computation process of K-NN is very difficult due to the procedure of classification which will utilize whole training samples. The K-NN algorithm will count the real distance and sort the distance of every one of the training data at each prediction [17, 20].

Another issue is combining the class's labels using a particular technique. K-NN algorithm has a set of samples (n labeled); in which n is the total data number included in the set that can be illustrated as:

D.A={da1, da2, …, dan}

D.A is the samples set, so each (da1, da2, da3,….., or dan) is independent sample which are differ from each other. The set of samples (n labeled) can easily be shown as:

D.A={da1, da2, da3 l C}

C is the class of the targeted value. As a result of the above example, we explain the mechanism of the KNN algorithm and how it used to classify the new unclassified data. The above example used for the purpose of classifying a small dataset.

4. DATA CLASSIFICATION

As the IoT is a huge communication environment, many data are traveling between IoT devices. It is really important to classify the data security level in the IoT environment. To explain the importance, as for instance if we have an organization with 150 devices connected in the IoT network. Obviously, these IoT devices are kept to communicate with each other by sending and receiving data. of course, these huge data need encryption technology to ensure their security goals. For sure, it is wasting time, effort and battery lifetime if consider that all data belong to the same level of security [21]. Thus, we classify the security level of data by implementing K-NN machine learning into four types as illustrated in Figure 1.

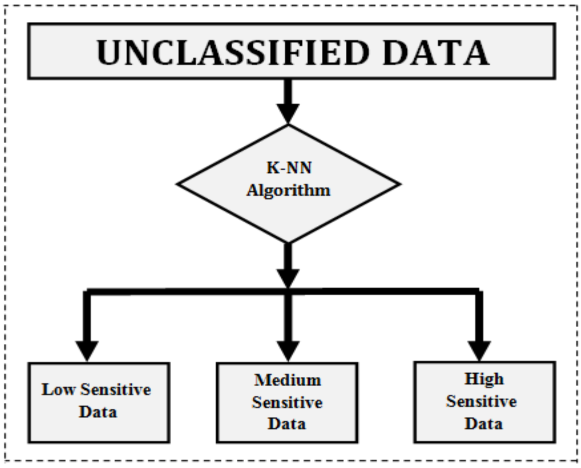

Figure 1. Classify data upon the security needs

4.1. Classify data into level of security

The universality of privacy translates as the proportion of the people who view a piece of information as to their privacy between all the people. On the other hand, the universality of privacy means the number of people who think they are assaulted when disclosed the information. Therefore, the universality of privacy can indicate the range of the parties that are involved in the assaulted of privacy [22]. The confidentiality of privacy indicates the privacy value of its secrecy degree.

a) Lowsensitive data:It is the type of data in which does not make a significant effect on the organization in case of a change or steal. The confidentiality and universality of privacy are low. Thus, the requirement of security is the lowest.

b) Medium sensitive data: It is the type of data in which has little effect on the organization in case of a change or steal. In this type of data, one of the following circumstances must be faced:

a. The confidentiality of privacy is high/medium, and the universality of privacy is medium or low.

b. The universality of privacy is high/medium, and the confidentiality of privacy is medium or low.

c. The universality and confidentiality of privacy is medium.

Thus, the requirement of security is higher than low sensitive data.

c) High sensitive data: It is the type of data in which has a significant effect on the organization in case of a change or steal. The confidentiality and universality of privacy are too high. Thus, the requirement of security is the highest.

Based on the confidentiality and universality of privacy, the Table 1 below has explained the decision of security classification, which it is consist into three level: low, medium, and high.

Table 1. The decision of Security classification based on the confidentiality and universality of privacy

		UNIVERSALITY OF PRIVACY		
		LOW	MEDIUM	HIGH
CONFIDENTIALITY OF PRIVACY	LOW	Low Sensitive Data	Medium Sensitive Data	High Sensitive Data
	MEDIUM	Medium Sensitive Data	Medium Sensitive Data	High Sensitive Data
	HIGH	High Sensitive Data	High Sensitive Data	High Sensitive Data

The organization owner is always worried about their data to be stolen, altered, modified or loss. These data may belong to personal data, financial transactions, business material, financial records, medical/health data, or government. In [23], the data classified based on security into three levels of security which are: (a) public data, (b) internal/privacy data, and (c) confidential data. Furthermore, in [24] there are four levels of sensitivity for data, which are (a) confidential/high, (b) confidential, (c) internal, and (d) public risk. Also, in [17], the data divided into two types which are (a) normal sensitive data, and (b) high sensitive data. Additionally, in [25] the data classified according to the privacy level into four levels which are (a) high security privacy,(b) medium security privacy, (c) basic security privacy, and (d) low security privacy.

5. FRAMEWORK DETAILS

We suggest that the data which are classified using the K-NN algorithm will use three different mechanisms of encryption, these mechanisms depend on level of sensitivity. So, after data pass through the K-NN algorithm the device will recognize the sensitivity level. So, the three cases will be described below:

5.1. Process at sending device

The sender will collect the data which is needed to be sent. Then, these data will digest by the K-NN algorithm. Then after, the K-NN algorithm will recognize the sensitivity level, here is one of three below decisions will be taken:

a) If the data belong to a low sensitive level, they will not apply any type of encryption mechanism, as the data are not having any type of impact in case of loss. Therefore, data will be sent plain to another device or to cloud if required to save or analysis.

b) If the data belong to a medium sensitive level, they will apply the advanced encryption standard (AES) encryption algorithm, as the data are having little effects in case of loss. Therefore, data will encrypt using AES algorithm. So the encrypted data will be sent along with its secret key to another device to cloud if required to save or analysis.

c) If the data belong to high sensitive level, they will apply a hybrid technique of RSA and AES algorithms. Using this technique, the data will be encrypted using the AES algorithm and the secret key will be encrypted using the RSA algorithm, as the data are having major impacts in case of loss. So the attacker cannot obtain the secret key of data even if they get encrypted data.

5.2. Process at receiving device

The receiver will gain the data from the sender. Then, these data will digest by K-NN algorithm. Then after, the K-NN algorithm will recognize the sensitivity level, here is one of three below decisions will be taken:

a) If the data belong to a low sensitive level, they will not apply any type of decryption mechanism, as the data did not encrypt in the sender's side.

b) If the data belong to a medium sensitive level, then the receiver will use the secret key of AES to decrypt the data.

c) If the data belong to a high sensitive level, then the receiver will use it's private key and RSA algorithm to decrypt the received encrypted secret key of AES, and then the secret key will be used with AES to obtain the plain text of data.

6. CONCLUSION

In the environment of IoT technology, all devices must be communicated and keep sending and receiving data. these data are transfer using communication channels such as Bluetooth, FRID and many more. Further, the transferred data between IoT devices may need to be sent outside the boundaries of its area to cloud through the internet channel. Whenever data transfer inside or outside the boundary, it will be vulnerable to attackers. The attacks which are inside the boundary of the IoT environment are different from the outside boundary. Thus, we have proposed a technique to classify the data according to its sensitivity to environment. The K-NN machine learning was implemented in this work to get the level of sensitivity whether its low, medium, or high. Then, according to the level, we propose the technique to safeguard the data during transmission. As future work, we are looking to implement the proposed work in a real IoT environment and evaluate it by doing a measure of CPU, Battery usage, and some other computing resources.

REFERENCES

[1] X. Lu, et al.,"Privacy information security classification study in internet of things,"*International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 162-165, 2014.

[2] L. Atzori, et al.,"The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787-2805, 2010.

[3] A. Kulkarni and D. Mukhopadhyay, "Internet of things based weather forecast monitoring system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 3, pp. 555-557, 2018.

[4] W. H. Dutton, "Putting things to work: social and policy challenges for the Internet of things," *Info*, vol. 16, no. 3, pp. 1-21, 2014.

[5] L. Edwards, "Privacy, security and data protection in smart cities: A critical EU law perspective," *European Data Protection Law Review*, vol. 2, no. 28, pp. 1-37, 2016.

[6] T. S. Gunawan, et al.,"Prototype design of smart home system using internet of things," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 107-115, 2017.

[7] M. U. Bokhari, et al.,"Security and privacy issues in cloud computing," *2016 3rd International Conference on IEEE Computing for Sustainable Global Development (INDIACom)*, pp. 896-900, 2016.

[8] N. C. Luong, et al.,“Data collection and wireless communication in Internet of Things (IoT) using economic analysis and pricing models: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2546-2590, 2016.

[9] I. A. T. Hashem, et al.,“The rise of “big data” on cloud computing: Review and open research issues,” *Information systems*, vol. 47,pp. 98-115, 2015.

[10] M. S. A. Mahmud, et al., “Internet of things based smart environmental monitoring for mushroom cultivation,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 3, pp. 847-852, 2018.

[11] C. P. Mayer,“Security and Privacy Challenges in the Internet of Things,”*Electronic Communications of the EASST*,vol. 17,2009.

[12] K. Aquilina,“Public Security Versus Privacy in Technology Law: A Balancing act?”*Computer Law & Security Review*, vol. 26, no. 2, pp. 130-143, 2010.

[13] C. M. Medaglia, et al.,“An overview of privacy and security issues in the Internet of things,”*The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, pp. 389-395, 2010.

[14] S. Bin and L. Yuan,“Privacy and Security in the Exploitation of Internet of Things,”*Journal of Dialectics of Nature*, vol. 2011, no.6, pp. 77-83, 2011.

[15] A. L. Jeeva, “Comparative analysis of performance efficiency and security measures of some encryption algorithms,” *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 3, pp. 3033-3037, 2012.

[16] Y. Sun, et al.,“Data security and privacy in cloud computing,” *International Journal of Distributed Sensor Networks*, vol. 2014, pp. 1-9, 2014.

[17] M. U. Bokhari, et al.,“Reducing the required time and power for data encryption and decryption using K-NN machine learning,” *IETE Journal of Research*, vol. 65, no. 2, pp. 227-235, 2019.

[18] N. Nodarakis, et al., “kdANN+: A rapid AkNN classifier for big data,”in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*, pp. 139-168, 2016.

[19] A. A. Abdulaziz, “Features Extraction Scheme for Behavioral Biometric Authentication Touchscreen Mobile Devices,” Doctoral dissertation, Universiti Teknologi Malaysia, 2016.

[20] Z. G. Liu, et al.,“Adaptive imputation of missing values for incomplete pattern classification,” *Pattern Recognition*, vol. 52, pp. 85-95, 2016.

[21] M. U. Bokhari and Q. M. Shallal, “Evaluation of Hybrid Encryption Technique to Secure Data during Transmission in Cloud Computing,” *International Journal of Computer Applications*, vol. 166, no. 4, pp. 25-28, 2017.

[22] M. U. Bokhari, et al.,“A Performance Analysis of Hybrid Technique using DES and RSA algorithms,”*2017 4th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2017.

[23] “Security Policies and Procedures,” Michigan Tech Information Technology.Available: http://www.mtu.edu/it/security/policies-procedures-guidelines.

[24] UTHSCSA Data Classification report, “Protection by data classification security standard,” Policy Ref: 5.8.21 Data Classification, 2006.

[25] X. Lu, et al.,“Privacy information security classification for internet of things based on internet data,” *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, 2015.