Journal of Innovative Engineering and Research (JIER)
Vol.- 3,Issue - 2, October 2020, pp. 1-11 (11 pages)

Privacy Preserving Data Gathering and Analysis for IoT Networks

MD ILYAS, RAJEEV RAGHUVANSHI, DR.DHANRAJ VERMA

Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore
Corresponding Author Email: er.mohd.ilyas @gmail.com

Abstract— The privacy preserving data mining is a technique which ensures the securely and privately mining of data in centralized or decentralized computation. The entire process involves collection of data, processing, and decision distribution. The combining data from various data sources (i.e. client, devices, servers, and other), increases dimensionality of data, and additionally results in communication, computational and storage overhead. In this context we need an efficient and effective technique which helps to achieve (1) a comparative study among cryptographic and noise based model (2) additionally an optional attribute selection algorithm is introduced which provides a way to reduce the dimension of data. (3) Improving the data utility after sanitization of data. Therefore two experimental models are proposed. First model is based on cryptographic method for achieving security and privacy. The AES and MD5 algorithm is used for cryptographic solution. Additionally for mining the encrypted data the Apriori algorithm and C4.5 decision tree is selected for experimentation. In the second experimental model we first prepare a noise model which accepts a range of noise level to introduce the noise over data. In next step the model compute the optional attributes to reduce the communication and computational over head. Finally the decision tree algorithm is employed to mine decision rules. This experiment is aimed to investigate the performance difference between noise based privacy preserving model and cryptographic PPDM technique. The comparative performance analysis is carried out in terms of accuracy, memory usages, and time consumption. The experimental evaluation demonstrates the noise based models are efficient then cryptographic models. The cryptographic models consume additional time for preparing cipher text. Secondly the dimensionality of data significantly impacts the performance of processing server as well as the communication network performance. Third, the controlled amount of noise can help to regulate data utility closer to actual data. Thus the proposed work is enhanced classical PPDM technique in order to compute decision rules accurately with less resource consumption in terms of network and communication.

Index Terms— PPDM, privacy and security, decision mining, data dimensions, resource utilization, noise based PPDM, cryptographic PPDM.

I. INTRODUCTION

The number of applications now in these days generating the data, such a huge data also includes a number of applications. On the other hand sometimes data are carrying the confidential and sensitive data. Therefore, privacy preserving techniques are required to secure the data in network as well as at servers. Thus the proposed work is motivated to design a lightweight, efficient and high utility model for privacy preserving of data. Let's consider a network with a large number of network nodes. All these nodes can collect, send, receive and combine their data. The collected data by nodes are moving towards the base station. The base stations are managing and mining the information to extract the knowledge. All these process in a system is known as privacy preserving data modeling. That scenario is beneficial for various applications such as IoT (Internet of Things) network, banking and finance, medical and health care, and others. In this area we found the following key issues:

1. During data aggregation, from the different parties, dimensions of data significantly increases, therefore a dimensionality reduction techniques is required
2. Which kind of security model can reduce the communication and computational overhead i.e. cryptographic or noise based approach

In this context we involved two data models first is based on association rules and decision tree in a privacy preserving environment. The aim of the experiment is to identify the suitable classifier for good performance of PPDM (Privacy Preserving Data Mining) system in cryptographic environment. Second is aimed to design a noise based model. That incorporates a dimensionality reduction algorithm and a noise inclusion algorithm for sensitizing the data. It also consist the comparative study among noise based technique and cryptographic technique. The proposed study supports various kinds of applications so the following objectives are proposed:

- **To implement a cryptographically secure PPDM model at the data supplier:** this module includes the development and design of PPDM model which is based on cryptographic security. The user encrypts their data by using a session key and communicates it for secure and private decision mining.
- **Handling data at the data aggregator for analyzing and distributing decisions:** the data aggregator can be connected with multiple data sources. The data aggregation becomes complex and resource consuming due to the data dimensionality. Therefore we need to implement a dimensionality reduction technique to reduce the data dimensions.
- **Recovery of data and decisions only at the authentic party:** the work also needed to design the data recovery technique at the authorized party. Therefore in cryptographic technique as well as in the noise based technique the data recovery system is prepared.

- **Reducing the dimension of data and improve the quality of data to ensure utility:** the proposed work include the study of data sanitization approaches that help to reduce the information losses and improve the data utility.
- **Preparing a combined framework for producing supervised learning based privacy preserving data modeling:** in PPDM environment for preparing the decisions for all the parties need to understand the utility of contributed part of data. Therefore the model includes the supervised learning techniques for proposed PPDM models.

II. RELATED WORK

The proposed work is motivated to design an efficient and accurate privacy preserving model which can support a large range of applications. In this context some key issues and challenges are discussed then their solutions are proposed.

A.Methodology

As the data is growing in a number of security flaws are also increases. In this work the data privacy and security is essential aspect of study. In a number of applications end user submit their personal and confidential information to grab the services. This data may used by the service providers to improve their business strategies. In this context the discloser of data at any level can harm the privacy of the end user. Thus, we need to design a secure, efficient and accurate way of data security and their processing. The following key issues are needed to be address:

1. data dimension can impact the performance
2. there are different PPDM models (i.e. cryptographic and noise based techniques) exist which kind of model is beneficial
3. it is also investigate how the noise or cryptography can impact on the classifiers performance

Thus the proposed work is developing two experimental models such as:

1. **Implementation of cryptographically secured PPDM:** in this phase a client and server model is prepared where the different clients are submitting their data using cryptographic algorithm. Additionally two data models are used for verifying the utility of data.
2. **Implementation of noise based PPDM:** in this phase a noise based data sanitization technique is used, additionally a dimension reduction algorithm is implemented.

B. Privacy Preserving Data model using Cryptography

This section discusses the cryptographic PPDM model. Additionally for achieving the objectives, thus a model is also demonstrated.

a. System overview

The work is intended to find an efficient and effective technique for mining privacy preserving decision rules [55]. In this environment, three actors are exist, first the data owner who have their own data, and for some essential task, submit their data to some organization or institution. That is necessary because without this information client does not get the required services. This data may involve some personal data, such as banking details, credit card information, and others. The second is an institution, which gathers various clients' data. These authorities can supply and use this data in favor of the client. But in some conditions, it is required to club their data with others for making business-oriented decisions. The individual part of data is not sufficient for making effective decisions.

But, the data discloser of end clients can affect the privacy and security. Therefore, before discloser of the data to any third party the sanitization of data required [56]. Finally, the third stack holder is the authority where the data is going to be disclosed. This authority is responsible for mining the data securely and privately, without disturbing the utility (application) of data [57]. Therefore, in order to deal with this situation the PPDM is required. In this context, two popular techniques namely association rule mining and decision tree-based rule mining techniques are used. The aim is carried out a comparative study for finding the superiority and key points for extension.

b. Proposed methodology

The proposed methodology includes three main layers as demonstrated in figure 1:

Connectivity and key generation: As we discussed the PPDM models includes the number of parties, who are contributing data. Therefore, when a party agreed for combine the data, then it is required to establish secure connection between the agreed parties and the server.

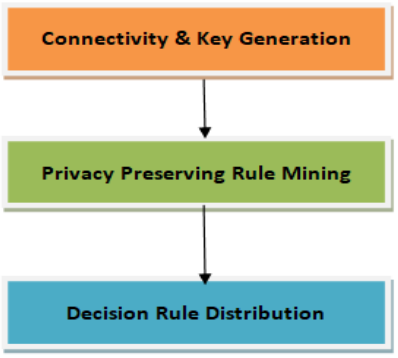

Figure 1: Layers of PPDM Model

In this context the following process is taken place as given in figure 2. The process is initiated when the data supplier (client) wants to combine the data with others, and start communication with the server.

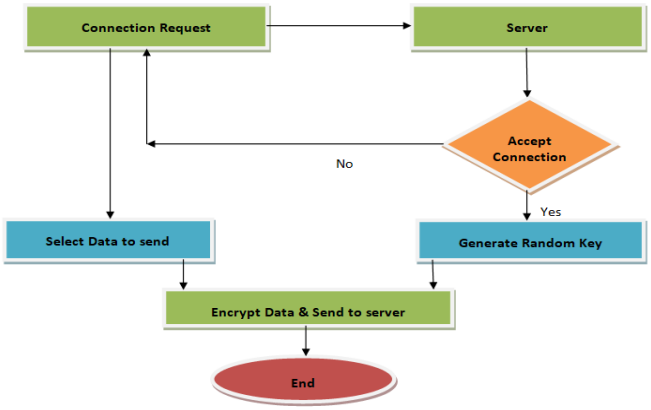

Figure 2: Connection process

Then, the client sends a connection request to the server. The server finds a connection request, then accepts the connection. After that, the server generates an 8 digit random number as the session key. The server sends this key to the client. The obtained key from server is used by the client to generate the encryption key. The encryption algorithm encrypts the input attributes and produced ciphered attributes. The cipher has been sent to the server for rule mining. In order to generate the cipher-text, the following process is used as demonstrated in figure 3. According to figure 3, the encryption technique accepts the data which is needed to be sent and the generated session key. The session key is passed over the MD5 algorithm, which produces the 128-bit hash code. This code is used as the encryption key. The AES algorithm is implemented for generating the cipher-text. The AES algorithm usages the 128 bit key and user input dataset attributes to encrypt the data.

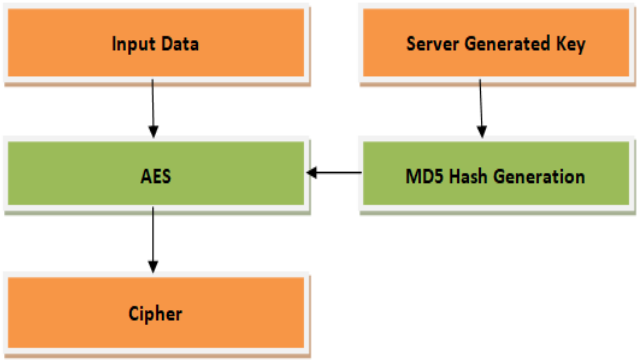

Figure 3: Encryption process

Privacy preserving Association Rule mining: The rules are mined on the basis of two popular algorithms, Apriori algorithm and C4.5 decision tree.

1. Apriori Algorithm

Apriori is a data mining algorithm for association rules mining based on candidate set generation. It is designed for transactional databases. In order to understand the transactional database, let a shop has a collection of items, purchased by consumers [58]:

- All subsets of a frequent item-set
- And, all the combinations of item-set, are considered

Prior to start, let us set the limits to half, for the support count.

Step 1: create a frequency table contains all the items with their occurrence based on the transactions.
Step 2: Those items are important that has the frequency higher than the support. The support value is selected by user, thus, items with less than support count are removed.
Step 3: Next we need to prepare the subsets of the items. It is required to consider the combinations AB is the same as BA. We create all the possible combinations with two items.
Step 4: The combinations are created with the two items we calculate the frequency in the transactions.
Step 5: after that, eliminate the item's pair which is lower than the support.
Step 6: Now create the pairs of the three items. Start the search over the transactions for finding the combinations with their frequencies higher than the support count. This process continued until all the combinations are not covered. The algorithm of the described process is given using table 1:

Table 1: Apriori Algorithm

Process **Apriori** (T,) {

;

$for\ (k = 2;\ L_{k-1}! = \emptyset; k++)\{$
$C_k = genrate\ Candidateset\ using\ L_{k-1}$

// It is Cartesian product and eliminating any size itemset that is below support count

$for\ each\ transaction\ t\ do\ \{$

#increment count of all candidates in that contain in t

= candidates in with minSupport

}

}

return

}

2. Decision tree C4.5 or J48

The second algorithm for decision rule mining is the C4.5 or J48 decision tree. That is an extension over ID3. Entropy and information gain is used to create data partitions. The attribute with higher information gain is used at a higher level. Algorithm continuously uses both steps to create sub-lists to create the tree. Thus information gain requires entropy first. To calculate the entropy, dataset contains two classes, T (True) and F (False). The entropy is measured for entire dataset D. For example, for binary classification the entropy E for dataset D is defined as:

$$E(D) = -T log_2 T - F\ log_2 F$$

To reduce the depth of tree, selection of the best possible attribute is required for branches. The attributes with minimum entropy will be selected. Thus information gain is required to drop with respect to an attribute during splitting. The information gain, Gain (E, A) for attribute A is,

$$Gain(E, A) = Entropy(s) - \sum_{n=1}^{v} \frac{E_v}{E} XEntropy(E_v)$$

The gain can be used to decide positions of attributes in the tree. The position of attribute depends on the two factors, first to create a small size tree, and to offer the required level of unfussiness. Algorithm returns the decision tree as learning consequences [59].

Now the process involved in this technique is discussed using figure 4. The connected clients are sending to the server which is aggregated over a common dataset. The cryptographic process transforms entire data attributes. The ciphered data is used with the apriori algorithm and C4.5 decision tree for generating the IF-THEN-ELSE rules. The rules are now ready to distribute among all the parties.

c. Decision Rule Distribution

The data distribution for all the clients not required any complex process for recovering the contributed part of data. Therefore the prepared rules are transmitted to all the connected clients. The clients are usages a similar key and algorithm for decrypting the rules. Using this blind decryption process only those parts of attributes are recovered which are contributed by the parties.

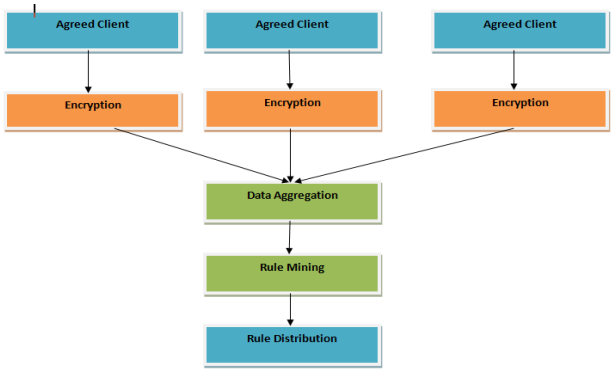

Figure 4: Rule generation process

C. Privacy Preserving Data modeling using Noise Based Technique

This section provides the details about the privacy preserving technique based on noise inclusion, which reduces dimensionality of data to preserve the server resources and communication overhead. In this context the previous method is enhanced for recovering the higher informative attributes. The decision tree algorithm is used in this experiment for computing the decision rules. Secondly, the data dimensions, which are increases during the aggregation of data, a new algorithm are introduced to reduce data dimensions and enhance the utility.

Over-fitted attribute removal

First we involve a technique to handle over fitting of the data. Due to over fitting, the decision tree is not learning with the entire data and cannot recover the fruitful outcomes. The table 2 is exploring the dataset attributes that can be used for learning.

Table 2: Reducing dimensions for over fitting

Input: Dataset D
Output: Reduced data R

Process:

A. $[row, col] = readDa$
B. $for(i = 1; i \leq$
   a. $for(j = 1; j \leq row;$
      i.
         1. $R[i,j].insert(Val(i,j))$
         $else\ if(Val(i,j) ==$
      ii. $R(i,j-1))$
         1. $coun$
         2. $R[i,j].insert(Val(i,j))$
      iii.
   b. End for
   c. $U = R.getUniqueC$
   d. $If(count$
      i. $R.Elim$
   e. if
      i. $R.Elim$
   f. End if
C. End for
D. Return R

Adding Noise on data

In place of encryption algorithm, a range based noisy attribute creation algorithm is used. However, the data set may have quantitative and/or qualitative attributes. To sanities the quantitative attributes we usages the data normalization. Thus, the min-max normalization is used which is given by the following equation.

$$newVal = \frac{current\ value - min}{max - min}$$

In addition, we provide a range for noise inclusion varies between 0-1. That random noise is introduced at contributing end. That input is given by NR. Thus new value can be given by:

$$newVal = newVal * NR$$

In Addition, for dealing with the qualitative attributes, we mutate each character of the attributes and create new values to replace existing values. The processes of the handling data is incorporated in table 3.

Table 3: Noise Adding algorithm

Input: input dataset from previous phase R, Noise Range NR, circular queue
Output: Noise added dataset ND

Process:

1. $[row, col] = R.GetDi$
2. $for(i = 1; i \leq$
   a. $for(j = 1; j \leq col$
      1. $if(R[j,i] == Neumeri$
         1. $newVal = \frac{R[j,i]-R.min}{R.max-R.mir}$
         2. $R[j,i] = newVal * NR$
      2. Else
         1. $Temp = R$
            $C[x] =$
         2. $Temp.GetC\square ar()$
            $for(k = 1; k \leq$
         3. $C.lengt\square; k++)$
            $C[k] = Q[k +$
            a. $NR]$
         4. $\{$
         5. $R[j,$
      3. End if
   b. $ND.$
   c. End for
3. End for
4. Return ND

After including noise to the dataset, the dataset attributes are transformed as noisy attributes. But the noise is distributed uniformly thus the utility of dataset is not being changed.

Measuring Data Utility

In order to measure data utility the correlation coefficient is suggested to be use. The correlation coefficient is denoted by r that told us relation between two vectors to define closeness of data. The closest value of r is 1. If r =1 or r = -1 then relationship perfectly defined. Data sets with values of r = zero show no relationship among them. Using this property we use it for electing the optional columns. To calculate the r we use the following eq.

$$r_{x,y} = \frac{\sum_{i=1}^{n} \quad d_x * d_y}{n\sqrt{\sum_{i=1}^{n} \quad d_x{}^2 * \sum_{i=1}^{n} \quad d_y{}^2}}$$

Where,

Deviation of variable

Deviation of variable

N = number of instances

Now the following algorithm is used for selection of optional attributes.

Table 4: Selecting optional attributes

Input: noise added data ND Output: Optional Data attributes O
Process: A. $[row, col] = ND.GetDir$ B. C. $for(i = 1; i \leq$ a. b. $CR_{V,C} = Calcula$ c. O D. End for E. $M = Calcula$ F. $for(j = 1; j \leq$ a. 1. $O_i.Mark(Optio$ b. End if G. End for H. Return O

The optional attributes demonstrate less likely hood with respect to the class labels. But it may be possible, the optional attributes will work better then locally involved attributes. And, it may also possible the optional attributes enhance the classification accuracy. Therefore, entire data is combined and used with correlation coefficient for finding relevancy among class labels and attributes. We found no change in their r value. Thus, if the attribute not has the strong relationship then we can reduce the attributes. Server accepts the data from all the clients, and organized according to the class labels. After that the decision tree is prepared using the newly created dataset. The decision tree produces the tree, which is further converts into the IF-THEN-ELSE rules. These rules are distributed to all the clients.

D.Data recovery

To recover the actual values at the client the similar technique is used as described in table 4.9. To implement the process we need to reverse the values which are generated. Therefore to recover the categorical values the sequences of characters are replaced with the existing values. Additionally to recover the numerical values, the following two steps are used.

$$newVal = \frac{Ot}{NR}$$

Where, Ot is the value which is received by the server and NR is the noise value included before, and the is min-max normalized data, we can recover the actual value using .

Initially we have,

$$newVal = \frac{current\ value - min}{max - min}$$

Thus,

$$newVal(max - min) = current\ value - min$$

And finally,

$$current\ value = newVal(max - min) + min$$

Thus if we have

$$\square = (max - min)$$

$$current\ value = newVal * \delta + min$$

Using the above given formulation the process involved is given in table 5.

Table 5: Data recovery algorithm

Input: Obtained rules R, Noise Range NR, circular queue , Output: Noise free data D
Process: 1. $[row, col] = R.GetDi$ 2. $for(i = 1; i \leq$ a. $for(j = 1; j \leq col$ 1. $if(R[j, i] == Neumeri$ 1. $newVal =$ $R[j, i] = newVal * +$ 2. min 2. Else 1. $Temp = R$ $C[x] =$ 2. $Temp.GetCar()$ $for(k = 1; k \leq$ 3. $C.lengt; k++)$ $C[k] = Q[k -$ a. $NR]$ 4. ‹ 5. $R[j,$ 3. End if b. D c. End for 3. End for 4. Return D

III. RESULTS ANALYSIS

The proposed work is aimed to design an improved privacy preserving data modeling system using two techniques that involves a cryptographic security and second is based on noise inclusion. This section provides performance analysis of both the models, and compares the performance of approaches.

A.Experimental Scenarios

In this work two experimental scenarios are involve both are explained in this section.

1. **Performance analysis of cryptographic PPDM Model:** in this experiment a cryptographic technique, based on AES and MD5 algorithm is used to encrypt data at client end. Additionally two kinds of mining techniques are used, namely Apriori algorithm and C4.5 Decision tree. The performance of the models is evaluated.
2. **Performance analysis of noise based model:** in this work two aims are involved, first a noise based model which is used to sanitize the data. Secondly an optional attribute selection technique is introduced to improve the memory and time resources. Using this approach we can also reduce the computational and communication overhead. The data mining technique is used to compute the decision rules. Finally the performances of the algorithms are measured.

B. Analysis of Cryptographic

The cryptography based privacy persevering rule mining technique is evaluated in this section.

a. Accuracy

In ML and data mining, accuracy of an algorithm is recognized as the number of decisions correctly made. Therefore, it is a ratio of total correct decisions and total samples for decision making. The following Eq. can be used.

$$accuracy\,(\%) = \frac{total\ correct\ decisions * 100}{total\ samples\ for\ decision\ making}$$

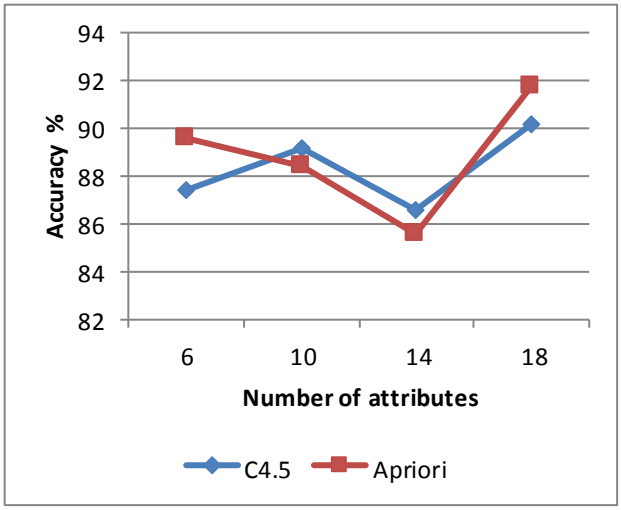

Figure 5: Accuracy

The accuracy of the privacy-preserving decision-making rule mining techniques using both the algorithms is reported in figure 5. The accuracy of techniques is given in the Y-axis, which is measured in terms of percentage (%). In addition, the X-axis of the algorithm shows the number of attributes. According to performance, the accuracy of both the algorithms is similar, but sometimes the performance of decision tree is better and sometimes the association rule mining based technique. However, the mean accuracy of the decision tree algorithm is higher as compared to association rule mining technique because of low ambiguity in decision tree-based rule mining.

b. Number of Rules

The performance of both the rule mining models namely association rule mining and the decision tree-based rule mining technique is described in this section by using the number of rule generation.

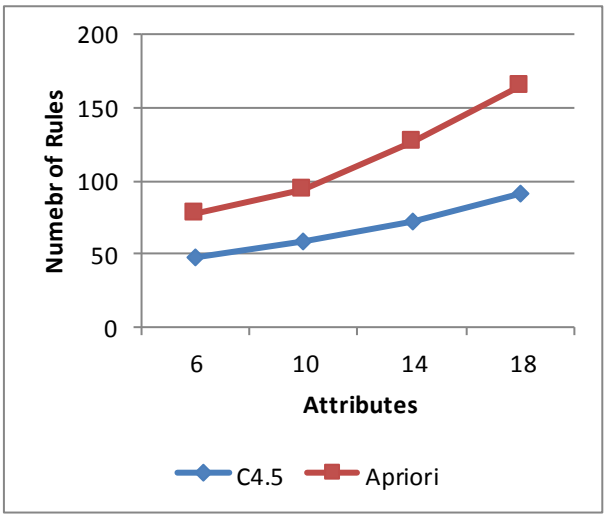

Figure 6: Number of rules

Figure 6 shows the number of rules generated using both the techniques in the Y-axis, and X-axis shows the number of attributes involved. According to the results, the association rules mining based technique generates large number of rules as compared to the decision tree-based technique. Therefore using the association rules consumes more numbers of cycles are higher than the decision tree-based technique.

c. Time complexity

The time complexity of an algorithm is the amount of time required to process the data according to the algorithm. That can be computed using the following Eq.

$$time\ consumed = Algorit\square m\ end\ time - start\ time$$

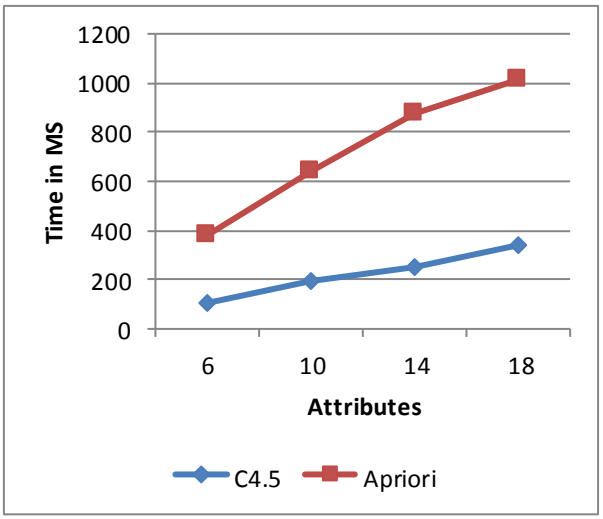

Figure 7: Time expenses

The time complexity of association rule-based technique and the decision tree-based technique is described in figure 7. The time expenses of the algorithms are measured in terms of milliseconds (MS). In this diagram, the Y-axis shows the time consumed and the X-axis shows the number of attributes utilized for experimentation. According to the results, the time consumption for the association rule mining technique is higher as compared to the decision tree-based technique.

d. Space complexity

The space complexity is also known as memory usages of the process, when a process initiated for execution the system assigns a fixed amount of main memory to the process. The difference of memory free and total assigned is known as the utilized memory. In JAVA that is computed using the following Eq.

$$memory\ usages = total\ assigned - free\ memory$$

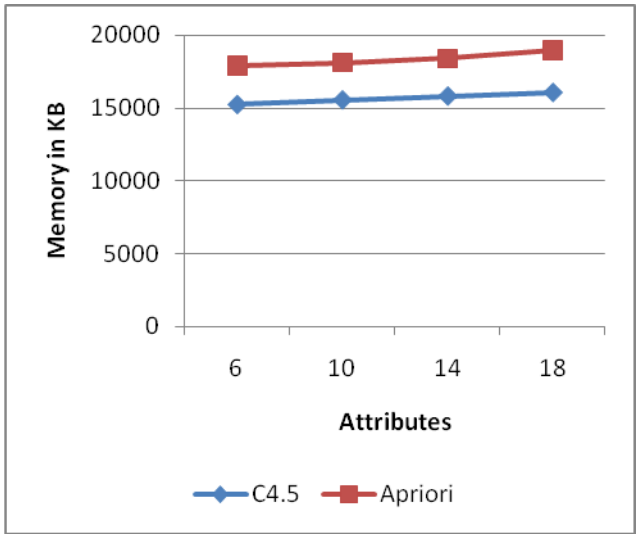

Figure 8: Memory usages

According to the obtained memory usages of association based and the decision tree-based rule mining techniques the results are demonstrated in figure 3.4. The memory of the systems is reported in the Y-axis and the X-axis shows the number of attributes involved. The experimental results demonstrate the memory usages of the techniques are increases with the number of attributes. Additionally, the memory usages of the association rule mining is significantly higher as compared to decision tree-based technique.

C.Noise Based Privacy preserving model

This section includes the comparison among proposed modified privacy persevering rule mining techniques and previously proposed technique. The similar parameters are used for comparison.

a. Aim of experiments

The proposed work is motivated to enhanced previously proposed PPDR model. That model includes encryption technique for sanitization of confidential data. That is replaced here with the noise based technique for hiding the sensitive and private information. In addition, to reduce the time and memory usages the utility of data is also measured and unreliable attributes were reduced. That may help to improve the computational cost. Therefore with the following aim the experiments are conduced.

1. To compare the modified technique with previously proposed model
2. To validate the utility of data after sanitization process adopted

b. Accuracy

The accuracy is known as correctly recognized ratio of samples by trained algorithm. Therefore, it is defined as total correct decisions and total samples. The following Eq. can be use.

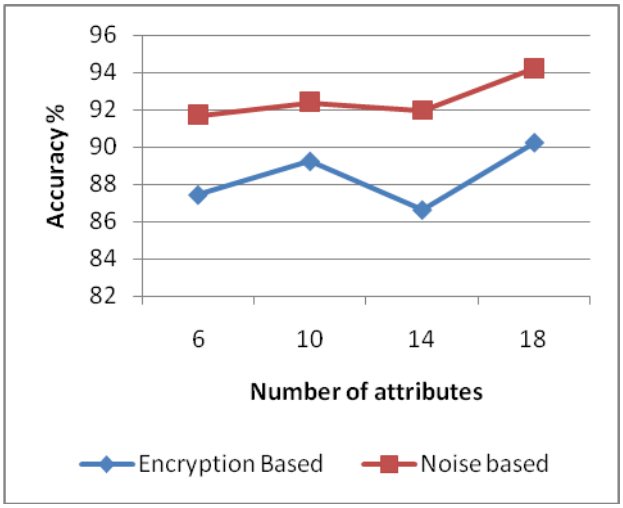

Figure 9: Accuracy

$$accuracy\,(\%) = \frac{total\ correct\ decisions * 100}{total\ samples\ for\ decision\ making}$$

Figure 9. shows accuracy of both the developed techniques. The Y-axis contains accuracy in percentage (%) and, X-axis shows the number of attributes as input. Accuracy of both the algorithms is similar. However, the mean accuracy of modified noise based PPDR algorithm is higher.

c. Space complexity

The space complexity of an algorithm is basically computed on the basis of process, when a process execution is initiated the system assigns an amount of memory to that process.

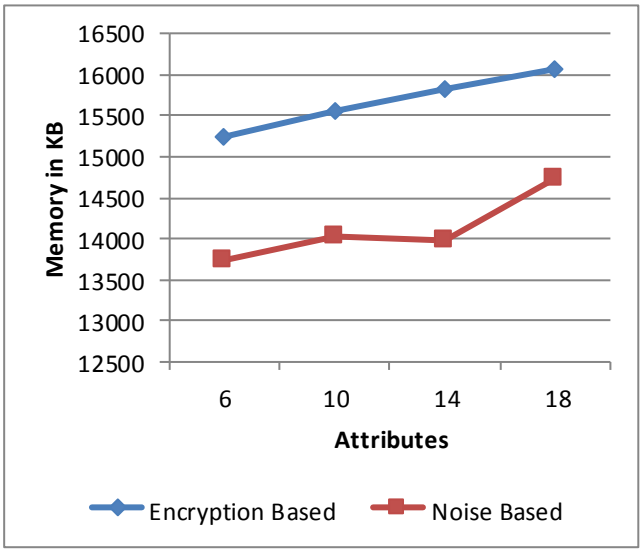

Figure 10: Memory usages

The amount of memory free from total assigned is known as memory usage. In JAVA it is computed as.

$$memory\ usages = total\ assigned - free\ memory$$

According to observation of memory usages of both the techniques is demonstrated in figure 10. The memory of the systems is given in Y-axis and the X-axis shows the attributes involved for rule mining. The results demonstrate the memory usages of PPDR technique are increases with the number of attributes. Additionally the encryption based technique is expensive as compared to noise based technique.

d. Time complexity

The time complexity or usages is the amount of time required to process the data using algorithm. That can be computed using the following Eq.

$$time\ consumed = Algorithm\ end\ time - start\ time$$

The time complexity of encryption based and noise based technique is reported in figure 11. The time consumption is noticed here in milliseconds (MS). The Y-axis of line graph shows the time consumed and the X-axis shows the number of attributes. The result says the time consumption of noise based technique is low as compared to encryption based technique. The line graph clearly demonstrates the gap between lines is increases with the size of attributes. Therefore the proposed technique can significantly for reducing the cost of algorithm execution.

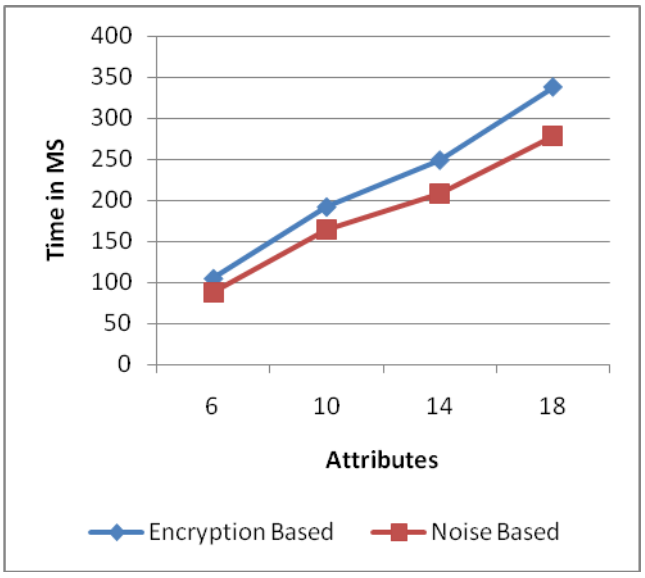

Figure 11: Time expenses

e. Number of Rules

The performance of both the PPDR model i.e. noise based technique and encryption based technique is described using number of rules. Figure 12 show the number of rules generated using both the techniques. The Y-axis includes the number of rules and the X-axis reports number of attributes involved in experiment.

According to the results, the noise based technique generates the similar number of rules as the encryption based technique. But the noise based technique mostly generates fewer rules as compared to encryption based technique.

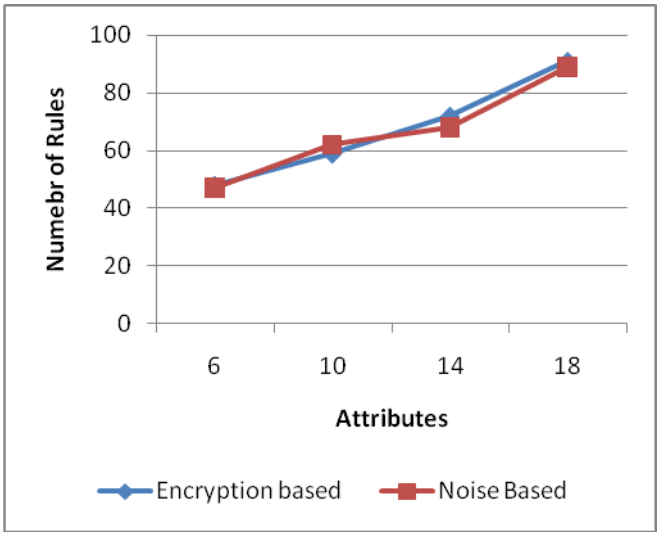

Figure 12: Number of rules

IV. CONCLUSION AND FUTURE WORK

This section involves the conclusion and future extension plan. The conclusion is made on the basis of experiments and theoretical investigation. The facts are reported as the conclusion and limitation of the work as future extension.

A.Conclusion

A number of applications exist where additional or delegated attributes are required for making precise decisions. In this context agreed parties are outsource their data. But, they are worried on discloser of data and decisions. Thus applications needed to sanitize and prevent the privacy loss of the end data owners. Therefore, we indented to improve the data modeling for the following:

1. Investigation and performance analysis of the cryptographic PPDM and noise based PPDM models
2. Identification of less essential attributes which can be reduce for improving the computational and communication overhead
3. Introducing a range based noise function that help to regulate the noise level on the training data and improves the data utility.

Therefore there are two different models for decision rule mining is proposed:

1. A cryptographic algorithm is implemented with the help of AES and MD5 algorithm. The MD5 usages user's session key to generate the cryptographic key and the AES algorithm is used to encrypt data in client end. Finally encrypted data is communicated to the server for decision rule mining. In order to mine decision rules C4.5 decision tree and Apriori algorithm is used. The experimental study demonstrates the C4.5 algorithm performs superior with higher data utility.
2. The next technique is focused on the decision tree algorithm to study the PPDM model. In order to handle the over fitting cases an outlier detection and removal algorithm is proposed. Using this algorithm user can reduce the attributes form the data. Further a noise mixture algorithm is proposed to sanitize the sensitive information. Then the data is being used for mining rules. But the data

utility is main concern. Therefore an enhancement also included to compute optional attributes. The optional attributes having less bounding with the target class labels. Therefore the data can be reduced at the client end to reduce the communication overhead and computational overhead. After that data attributes are transmitted to the server where the data is processed using decision tree algorithm. The rules are recovered and distributed to all the parties.

The proposed models (i.e. encryption based and noise based model) are developed using the JAVA. Moreover, to maintain and preserve the performance the MySQL Database is used. The system is evaluated using different parameters, and their summary is given in table 6.

Table 6: Performance summary

S. No.	Parameters	Encryption based	Noise based
1	Accuracy	Low	High
2	No. of rules	Higher	Low
3	Memory usages	Higher	Low
4	Time expenses	Higher	Low

According to the experimental results, both the models are enhancing the performance of existing system in terms of resource consumption and utility of the data. The improvement is noticed for data utility, time consumption and the memory requirements.

B. Future Work

The main aim of the proposed work is to improve applications of Privacy and maintaining data utility. The experimental study according to the aim is achieved successful using two models. In near future the following work is proposed.

1. Most of models are developed for either the rule based mining (association rules and decision rules). Additionally the supervised learning approaches are used. In near future the work is focused on unsupervised learning process
2. The current work is focused on the data discloser cases in near future the model is developed which works for human error based data discloser
3. Employing the proposed methodology under the real world application
4. Improve the model to accept heterogynous data types.

REFERENCES

1. C. Hu, J. Luo, Y. Pu, J. Yu, R. Zhao, H. Huang, T. Xiang, “An Efficient Privacy-Preserving Data Aggregation Scheme for IoT”, Springer International Publishing AG, part of Springer Nature 2018, WASA 2018, LNCS 10874, pp. 164–176, 2018.
2. S. Sharma, K. Chen, A. Sheth, “Towards Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems”, IEEE Internet Computing, March-April 2018.
3. E. Toch, B. Lerner, E. B. Zion, I. B. Gal, “Analyzing large-scale human mobility data: a survey of machine learning methods and applications”, Knowl Inf. Syst.
4. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15, 104–116, 2017.
5. J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, “A survey of machine learning for big data processing”, EURASIP Journal on Advances in Signal Processing, 67, 2016.
6. C. W. Lin, T. P. Hong, H. C. Hsu, “Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining”, Hindawi Publishing Corporation Scientific World Journal Volume, Article ID 235837, 12, 2014.
7. P. S. Rao, S. Satyanarayana, “Privacy preserving data publishing based on sensitivity in context of Big Data using Hive”, J Big Data 5, 20, 2018.
8. O. Tene, J. Polonetsky, “Big Data for All: Privacy and User Control in the Age of Analytics”, 11 Nw. J. Tech. & Intell. Prop., 239 , 2013.
9. R. Mendes, J. P. Vilela, “Privacy-Preserving Data Mining: Methods, Metrics, and Applications”, Vol. 5, IEEE, 2017.
10. Y. A. A. S. Aldeen, M. Salleh, M. A. Razzaque, “A comprehensive review on privacy preserving data mining”, Springer Plus 4, 694, 2015.
11. K. Xu, H. Yue, L. Guo, Y. Guo, Y. Fang, “Privacy-preserving Machine Learning Algorithms for Big Data Systems”, 2015 IEEE 35th International Conference on Distributed Computing Systems, 1063-6927, European Union, 2015.
12. I. San, N. At, I. Yakut, H. Polat, “Efficient paillier cryptoprocessor for privacy-preserving data mining”, Security and Communication Networks, Wiley Online Library, 9, 1535–1546, 2016.
13. Z. Gheid, Y. Challal, “Efficient and Privacy-Preserving k-Means Clustering for Big Data Mining”, IEEE TristCom, Tianjin, China pp.791 - 798, Aug 2016.
14. R. Lu, K. Heung, A. H. Lashkari, A. A. Ghorbani, “A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT”, Special Section on Security and Privacy in Applications and Services for Future Internet of Things, Vol. 5, IEEE, 2017.
15. Y. Kokkinos, K. G. Margaritis, “Confidence ratio Affinity Propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining”, Neurocomputing, vol. 150, pp. 513−528, 2015.
16. S. Sharma, K. Chen, A. Sheth, “Towards Practical Privacy-Preserving Analytics for IoT and Cloud Based Healthcare Systems”, IEEE Internet Computing, March-April 2018.
17. Y. Li, C. Bai, C. K. Reddy, “A Distributed Ensemble Approach for Mining Healthcare Data under Privacy Constraints”, Inf Sci (Ny). February 10, 245–259, 2016.
18. H. Hammami, H. Brahmi, I. Brahmi, S. B. Yahia, “Using Homomorphic Encryption to Compute Privacy Preserving Data Mining in a Cloud Computing Environment”, EMCIS LNBIP 299, pp. 397–413, Springer International , 2017.
19. K. Birman, M. Jelasity, R. Kleinberg, E. Tremel, “Building a Secure and Privacy-Preserving Smart Grid”, ACM SIGOPS OSR, 49, 1, pp131–136.

20. N. Domadiya, U. P. Rao, "Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases", Sådhanå 43:127 Indian Academy of Sciences, 2018.
21. W. Gan, J. C. W. Lin, H. C. Chao, S. L. Wang, P. S. Yu, "Privacy Preserving Utility Mining: A Survey", arXiv:1811, 07389, 2018.
22. J. C. W. Lin, W. Gan, P. F. Viger, L. Yang, Q. Liu, J. Frnda, L. Sevcik, M. Voznak, "High utility-itemset mining and privacy-preserving utility mining", Perspectives in Science 7, 74—80, 2016.
23. G. Kalyani, M. V. P. Chandra Sekhara Rao and B. Janakiramaiah, "Privacy-Preserving Association Rule Mining Using Binary TLBO for Data Sharing in Retail Business Collaboration", Advances in Intelligent Systems and Computing 515, Nature Singapore Pte Ltd. 2017.
24. B. Abidi, S. B. Yahia, C. Perera, "Hybrid Micro-aggregation for Privacy-Preserving Data Mining", arXiv:1812.01790v1, 4, 2018.
25. C. Y. Lin, Y. H. Kao, W. B. Lee and R. C. Chen, "An efficient reversible privacy- preserving data mining technology over data streams", SpringerPlus 5:1407, DOI 10.1186/s40064-016-3095-3, 2016.
26. N. Zhang, W. Zhao, "Privacy-Preserving Data Mining Systems", Published by the IEEE Computer Society 0018-9162/07/$25.00 © 2007 IEEE.
27. B. K. Pandya, U. K. Singh, K. Dixit, "A Study of Projection based Multiplicative Data Perturbation for Privacy Preserving Data Mining", International Journal of Application or Innovation in Engineering & Management, Vol. 3, Issue 11, Nov. 2014.
28. C. W. Lin, T. P. Hong, H. C. Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 235837, 12 pages,
29. S. Parmar, Mrs. P. Gupta, Ms. P. Sharma, "A Comparative Study and Literature Survey on Privacy Preserving Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 4, pg.480 – 486, April 2015.
30. S. Patel, K. R. Amin, "Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique", International Journal of Computer Science & Engineering Technology, Vol. 4 No. 05 May 2013.
31. T. Jahan, G. Narasimha, V. G. Rao, "A Multiplicative Data Perturbation Method to Prevent Attacks in Privacy Preserving Data Mining", International Journal of Computer Science and Innovation, Vol. 2016, no. 1, pp. 45-51, ISSN: 2458-6528
32. H. Zakerzadeh, C. C. Aggrawal, K. Barker, "Towards Breaking the Curse of Dimensionality for High-Dimensional Privacy: An Extended Version", arXiv:1401.1174v1 [cs.DB] 6 Jan 2014.
33. R. Heckel, M. Tschannen, H. B¨olcskei, "Subspace clustering of dimensionality-reduced data", arXiv:1404.6818v1 [cs.IT] 27 Apr 2014.
34. A. Aristodimou, A. Antoniades, C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection", Healthcare Technology Letters, 2016, Vol. 3, Iss. 1, pp. 16–21
35. Y. Wang, Y. X. Wang, A. Singh, "A Deterministic Analysis of Noisy Sparse Subspace Clustering for Dimensionality-reduced Data", Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s)
36. M. J. Zaki, W. MeiraJr, "Data Mining and Analysis Fundamental Concepts and Algorithms", Cambridge University Press Hardback, 2014 [Book]
37. M. Goebel, L. Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", ACM, 1999
38. N. adhabPadhy, Dr. P. Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), PP. 43-58 Vol.2, No.3, June 2012
39. N. Sundaravaradan, M. Marwah, A. Shah, N. Ramakrishnan, "Data mining approaches for life cycle assessment", In Sustainable Systems and Technology (ISSST), 2011 IEEE International Symposium on, pp. 1-6. IEEE, 2011
40. F. Gorunescu, "Data Mining: Concepts, Models, and Techniques", Springer, 2011
41. Zhao, Yijun. "Data mining techniques." (2015)
42. Ms. B. D. Pallavi, Prof. M. M. Waghmare, "Privacy-Preserving in Outsourced Transaction Databases from Association Rules Mining", International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
43. V. R. Redekar, Dr. K. N. Honwadkar, "Privacy-Preserving Mining of Association Rules in Cloud", International Journal of Science and Research (IJSR), Volume 3 Issue 11, November 2014
44. D. Tiwari, R. G. Tiwari, "A Survey on Privacy Preserving Data Mining Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 5, Ver. III (Sep. – Oct. 2015), PP. 60-64
45. J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in proceedings of 11th IEEE International Conference on Data Mining Workshops, IEEE 2011.
46. T. Jahan, G. Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy", in proceedings of IEEE 2012
47. S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012
48. A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in proceedings of International Symposium on Computer Science and Society, IEEE 2011
49. Kantarcioglu, M., Clifton, C.: Privacy preserving distributed mining of association rules on horizontally partitioned data. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 24–31, 2002.
50. P. Samarati. Protecting respondent's privacy in micro data release, TKDE, 13(6):1010–1027, 2001.

51. Vaidya, J., Clifton, C., Privacy preserving association rule mining in vertically partitioned data. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639–644. ACM Press 2002.
52. Brickell, Justin, and Vitaly Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008.
53. T. Pawar, Prof. S. Kamalapur, "A Survey on Privacy Preserving Decision Tree Classifier", International Journal of Engineering Research and Applications, Vol. 2, Issue 6, , pp.843-847, November- December 2012.
54. I. Ray, T. C. Ong, I. Ray, M. G. Kahn, "Applying Attribute Based Access Control for Privacy Preserving Health Data Disclosure", 978-1-5090-2455-1/16/$31.00 ©2016 IEEE, 2016.
55. L. Urquhart, N. Sailaja, D. McAuley, "Realising the right to data portability for the domestic Internet of things", Pers Ubiquit Comput 22:317–332, 2018.
56. K. K. Mishra, R. Kaul, "Audit Trail Based on Process Mining and Log", International Journal of Recent Development in Engineering and Technology, Volume 1, Issue 1, Oct 2013.