Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

A Privacy Preserving Framework To Mine Horizontal And Vertical Partitioned Data

Kamlesh Ahuja , Dr. Navneet Sharma

Department of Computer Science & IT, IIS (deemed to be University), Jaipur Rajasthan 302020

Abstract

Mining data in a central repository and maintaining security as well as data utility in the heterogeneous database is a complex task. In these conditions, we select the use of Privacy-Preserving Data Mining (PPDM). But most of the existing models are focused on vertically partitioned data. In order to address this issue, we proposed a Light Weight and End to End Cloud-based (LWE2EC) Privacy-Preserving Data Mining (PPDM) model. The objective is to offer efficient, secure, private, data utility, decision mining, and scalability for processing data in both formats i.e. vertically and horizontally partitioned. Thus both types of data organizations are studied. Additionally, the recent contributions in PPDM are also explored. The model offers the user to upload their data, encrypt it, and then collaborate with other data sources. This process generates a new dataset that contains multiple parties of data in an encrypted format and is associated on the basis of class labels. This dataset is further encoded using an attribute mapping process, then used with the data mining algorithm. On the other hand for secure and recoverable delivery of consequences of PPDM, to the collaborated parties the data is again transformed based on reverse attribute mapping. At the end of data contributors, the decision rules have appeared in an encrypted format where the user decrypts only those parts of decisions that are contributed by their data. This process increases the time and memory utilization and is only suitable for vertically partitioned data. Therefore, two modifications are applied, first, the client end is enabled to identify less informative attributes, by which data contributors can validate their data, and only send relevant information. This process also reduces the communication overhead and computational overhead. Secondly, the provision is made to accept the horizontally partitioned data. This process is also simulated using the relevant GUI. Both the models are compared with a baseline model. Based on the comparison in terms of accuracy, error rate, memory usages, and time, we identify that models are closer to the required data utility in both environments.

Keywords: PPDM, vertical partitioned data base, horizontal partitioned database, cryptographic security, SHA1, AES encryption, data utility, data recovery, decision mining.

1. Introduction

Data mining is a helping hand of new generation applications, which involve the use of machine intelligence and computational technologies as a giant infrastructure known as the cloud. With the technological growth new security and privacy issues are rising. In this work, we are investigating the security and privacy issues in a cloud for mining multi-party data without data leakage. However, there are a number of business domains, which are dependent on each other. For instance, the tour and travel industry depends on cab service, restaurants, and hotels, and connected with each other. In addition, there are other stakeholders are also present. In this condition, if all the stakeholders are want to achieve a common goal. Business analytics may helpful. In this approach using Machine Learning (ML) algorithms, we mine patterns to make data-centric decisions. But in these scenarios, the stakeholders are worried about the data confidentiality and privacy. Therefore, the parties are not agreed to disclose actual data. Additionally, after applying the ML algorithms, all the parties want to get their outcomes without disclosing their attributes and values.

In this context, security and privacy may depend on network and mining services providers. Suppose what happens if a server is compromised, thus to deal with such concern we cannot trust anyone. Therefore, this paper is motivated to develop a secure multi-party privacy-preserving data mining model name LWE2EC (Light Weight and End to End Cloud-based)-PPDM (Privacy-preserving data mining), to prepare an efficient data sanitization process. That helps to control privacy from source to the end. Additionally, we also work to mine both kinds of data formats i.e. horizontal as well as vertically partitioned data using a common framework. The framework contains three phases i.e. data pre-processing, pattern learning, and data publishing. Therefore, to maintain security and privacy the following objectives are proposed:

1. **To investigate cloud-based PPDM:** in this phase, the relevant literature to PPDM has been explored, and learns about the various data mining, security, and cloud computing concepts.
2. **To design an enhanced data gathering and mining environment:** we utilize the literature's experience in designing a lightweight, efficient, and accurate PPDM framework.
3. **Making ease in mining and publishing of the mining outcomes:** the work is providing end to end security and reduces the privacy disclosure risk.
4. **Performance analysis of the LWE2EC model:** the LWE2EC models are compared against the baseline model for measuring the fluctuation among the baseline model and proposed model.

2. Background

PPDM considers data mining to be uncovered the essential insights securely. The aim is twofold first, sensitive data like identifiers, names, addresses, etc, to be altered, for maintaining security and privacy. Second, sensitive data can be mined by data mining. Thus, security happens in two criteria: clients' own data and data aggregate on a server [2].

- **Individual privacy preservation:** The objective of data security is the insurance of data. Data is recognizable in the event that tends to be connected, directly or indirectly way, to identify an individual. Thus, the attributes of people are keeping private and shielded from exposure.
- **Collective privacy preservation:** Securing individual information may not be sufficient, we need to ensure against entire learning sensitive data. The aim is to assure the security of data when aggregated. The objective is to secure databases, in which we collect entire data, additionally secure recognizable data, and, also those data which may help to distinguish a person.
- **Limitations of PPDM:** PPDM doesn't mean immaculate security. The SMC calculation won't uncover the delicate information, yet the data mining result will empower to appraise the estimation of the sensitive data. It isn't that the SMC was "broken", however, that the outcome itself disregards security.

PPDM has a significant number of contributions claiming to design a robust technique. After review, we conclude some key challenges to be addressed:

- **Current approaches are not supporting re-query after data manipulation:** In a PPDM multiple parties are involved, but no one wants to disclose the sensitive records. Therefore handling all the security and privacy requirements is a complicated task [3].
- **Nature of data aggregator:** In PPDM scenarios a centralized database is required. So, required a secure and trustworthy place, but practically there is no fully trusted data aggregator [4].
- **Data formats vertically and horizontally partitioned are not feasible in a common framework:** Not only the security and privacy is a key concern but managing the data and association is also a challenge in PPDM [5].
- **Downgrading the learning performance:** In PPDM for the data mining ML techniques are used. The quality of data played an important role in learning. But to maintain the security of data the noise or encryption policies are applied. That impacts the performance of mining algorithms [6].
- **There is no provision to validate privacy after publishing a dataset:** In PPDM, a number of different data sources are participating to create a single dataset. Here the identification of meaningful attributes and the higher dimensions of data is a challenge. Additionally, after mining, results in distribution are also a complicated task [7].
- **After data publishing data leakage issues:** In a PPDM environment some of the applications are available where the publishing of data sets for third party use is required. Identification of sensitive and private information and disclosure is required. Handling of such attributes remains a research direction [8]..

3. Literature Survey

This section includes the study of recent contributions in order to improve PPDM models. Therefore some essential and noteworthy articles are collected and reported.

Data-as-a-Service (DaaS) enables data providers to integrate their data on demand. This involves some challenges like a mash-up data from multiple sources to resolve consumers' requests might reveal sensitive information and compromise privacy. M. Arafati et al [9] give

a cloud-based structure to privacy-preserving that empowers secure effort between Daa S suppliers to produce a secure dataset. Investigations show that Daa S is versatile, proficient and adequately fulfill the security and mining necessities. L. Li et al [10] center on PPDM on vertically partitioned databases. The data owners wish to get familiar with the visit of item sets from an aggregated dataset and uncover data as conceivable to other proprietors. To guarantee security Homomorphic Encryption (HE) and a secure correlation coefficient are used. And propose a cloud-based frequent itemset mining. The solution is intended for those databases that permit numbers of data owners to share data. The model release less private information and gives 3 to 5 degrees higher accuracy. They show that the run time is just one request higher than non-PPDM methods.

C. W. Lin et al [11], proposed a HMAU calculation to conceal touchy item sets through cancellation. The exchange with the maximal proportion of touchy to non-delicate is chosen to erase. The impacts of concealing disappointments, missing item sets, and fake item sets are thought of. The loads are alloted as the significance, as indicated by the necessity. Tests show the calculation in execution time, various erased exchanges, and various incidental effects. Then again, measurements show the quantity of information spill has been becoming because of human slip-ups. The current arrangements are restricted. Subsequently a methodology is needed for location. X. Shu et al [12] present security protecting DLD to settle the issue utilizing an uncommon arrangement of information digests. The benefit is that it empowers the information proprietor to designate the location to a semi-genuine supplier. The outcomes show that the technique can uphold exact recognition.

For the security of protection, touchy information have been scrambled, which makes data set use a test. J. Li et al [13] propose L-Enc DB, lightweight encryption, which keeps the information base construction and supports effective SQL inquiries. That can be utilized to encode a wide range of strings. The investigation exhibits that it is effective and secure. To work on the proficiency of large information highlight learning, Q. Zhang et al [14] proposes a protection safeguarding profound calculation model. To secure the private information, the model uses BGV encryption and utilizes cloud servers to play out the back-spread for profound learning. The scheme uses sigmoid function to support the secure computation. In the scheme, only the encryption and decryption are performed by the client. Results show that the scheme is improved by approximately 2.5 times in the training.

Table 1 Review Insights

Ref. no	Research type	Data and methods	Key insights
[9]	Research	Privacy preserving Daa S to empower secure and coordinated effort between Daa S suppliers.	Versatile, proficient and adequately fulfill the security and mining needs.
[10]	Research	HE and a secure correlation coefficient is used	For databases that permit multiple data owners to share data safely

[11]	Research	Hiding-Missing-Artificial Utility algorithm to hide sensitive item sets.	Effects of hiding failures, missing item sets, and artificial item sets
[12]	Research	Solutions of data leaks by human are limited, and challenging. Present a privacy preserving DLD using a set of data digests.	Enables data owner to delegate the detection to a semi-honest provider
[13]	Research	For privacy, we have encrypted data before outsource, which makes data utilization challenging. Propose L-Enc DB, an encryption, which keeps database structure, and supports efficient SQL-based queries.	A new format-preserving encryption scheme is constructed, which can be used to encrypt all types of strings.
[14]	Research	To improve the efficiency of big data feature learning, proposes a privacy preserving deep computation model by offloading the expensive operations.	BGV to encrypt private data and cloud to perform BPN algorithm for deep training
[15]	Research	To protect privacy of the outsourced data and association rules mining, k-anonymity, k-support, and k-privacy have proposed by perturb. The solutions are built on El Gamal.	To reduce the possibility of servers are compromised, user can select servers from different providers.
[16]	Research	Achieve PPDM where data are distributed and shared. Utilize data locality of Hadoop and limit number of cryptographic operations.	The scheme is secure in the semi-honest model.
[17]	Research	Introduce CGs then transfer CGs into a linear form by modification and mapping. Then use multi-keyword ranked search and raise PRSCG and PRSCG-TF.	Most existing efficient and reliable cipher-text search schemes are based on keywords.
[18]	Research	Method for generating a privacy-preserving heat map with user diversity (ppDIV), in which the density of trajectories, and diversity of users, is taken.	It was introduced as a pre-processing step following the principle of k-Anonymity
[19]	Research	Multi-objective algorithm to find optimal sanitization attributes. The GMPSO uses pre-large to speed up the process, and reduce multiple database scans.	Existing solutions are based on single-objective. GMPSO achieves better effects, and speed
[20]	Research	MPC framework for large-scale data mining. Priv Py combines easy-to-use and flexible interface with secret-sharing-based MPC backend.	It can support many real-world ML algorithms and large datasets with minimal algorithm porting effort.

[21]	Research	Privacy-Preserving and Security Mining Framework (PPSF), focuses on PPDM and data security	Offers algorithms for: data anonymity, PPDM, and (3) PPUM
[22]	Research	Architecture for PPDM based on MPC and secures sums. Two different protocols are proposed and measures failure probability is analytically modeled.	Privacy degree, communication cost and computational complexity are also characterized.
[23]	Review	The analysis of PPDM algorithms should consider the effects of these algorithms in mining the results as well as in preserving privacy.	The success of PPDM is measured using data utility, uncertainty level, anonymization, and randomization.
[24]	Research	Privacy-preserving data aggregation is one of typical fog applications. Existing solution only support homogeneous devices, and not aggregate hybrid devices'.	LPDA considers Paillier encryption, Chinese Remainder Theorem, one-way hash chain to aggregate hybrid devices' data and filter false data.
[25]	Research	However, HE can protect the data in theory, it has not been well utilized because it is too slow, especially multiplication.	Design logic of atomic operations in encrypted and apply logic to well known algorithms and also analyze the execution time of algorithms.

In PPDM, to protect the privacy of data and association rules, k-anonymity, k-support, and k-privacy have been proposed by X. Yi et al [15] to perturb the data. These techniques are expensive. The author considers a scenario where a data owner encrypts data and stores it in the cloud. To mine rules, the user outsources the task to "semi-honest" servers. They provide solutions to protect data privacy. The solutions are built on the distributed El Gamal, and to reduce the possibility of compromised servers; the user can select servers from different providers. Sometimes big data may involve multiple organizations that may have a different privacy policy, and may not share data publicly while joint data processing may be a must. K. Xu et al [16] propose to achieve privacy-preserving ML where the training data are distributed and shared. They utilize the data locality of Hadoop and the limited number of cryptographic operations is performed.

Most existing efficient and reliable cipher-text search schemes are based on keywords. Z. Fu et al [17] propose a content-aware search, which can make semantic search smart. First, introduce CGs then, present two schemes. They transfer CGs into a linear form with some modification and mapping. Second, employ the multi-keyword ranked search against two threat models and raise PRSCG and PRSCG-TF to solve the problem. J. Oksanen et al [18] have

developed a method for generating a privacy-preserving heat map with user diversity, the density of trajectories, and diversity of users, are taken. The method is applied to public cycling workouts and compared with privacy-preserving kernel density estimation on the density of trajectories and privacy-preserving user count calculation. It was introduced as a pre-processing step using k-Anonymity.

Information sterilization is a way of bothering a data set and conceal touchy data. Numerous calculations have been examined, albeit most depend on single-target techniques to find the applicant exchanges/things for disinfection. T. Y. Wu et al [19] present a multi-target calculation utilizing a matrix-based strategy. The GMPSO utilizes two systems for refreshing gbest and pbest. In addition, the pre-huge is adjusted to accelerate the cycle and lessens different information base sweeps. From the GMPSO, numerous Pareto arrangements instead of single-target calculations can be inferred. Also, the results of disinfection can be diminished. Trials have shown that accomplishes preferable impacts over the past calculation and can accelerate the calculation. Y. Li et al [20] present a MPC structure. Priv Py consolidates a simple and adaptable interface with a mystery sharing-based MPC backend. With fundamental information types and activities, and programmed code-revising is utilized. Show that it can uphold true ML calculations and huge datasets.

J. C. W. Lin et al [21] present a PPSF. It is an open-source library, which offers calculations for information secrecy, PPDM, and PPUM. PPSF has an easy-to-use interface for running calculations and showing the outcomes and is a functioning undertaking with customary deliveries. M. L. Merani et al [22] propose a design for PPDM dependent on MPC and secure totals. While conventional MPC chips away at less number assembly peers without a confided in substance, the review gives an answer that includes all information sources in the total interaction. A huge scope situation is inspected and the likelihood that information becomes out of reach is thought of. It is examined, as it might be provoked by intermittent network connectivity or sudden user mobility. For reliability, data sources are organized in multiple sets, which work on aggregation. Two different protocols are proposed and the measures failure probability. The privacy degree, communication cost, and computational complexity are characterized. Finally, the protocols are applied to specific use cases, to demonstrate their feasibility. According to M. M. Siraj et al [23], online application brings privacy threats to the data. There was been growing concern of violating individual privacy. The analysis of PPDM algorithms should consider the effects of these algorithms in mining results and in securing privacy. The success of PPDM algorithms is measured in terms of data utility, level of uncertainty, data anonymization, and data randomization.

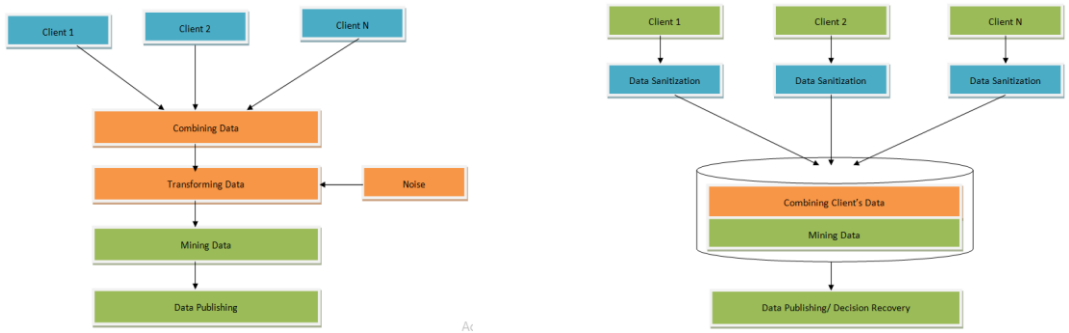
Privacy-preserving data aggregation is one of the typical fog applications. Most of the existing techniques only support homogeneous devices, and cannot aggregate hybrid devices' data. R. Lu et al [24] present a lightweight privacy-preserving data aggregation. The LPDA is characterized by Paillier encryption, Chinese Remainder Theorem, and one-way hash chain techniques to aggregate hybrid IoT devices' data, and also early filter false data. Analysis shows LPDA is secure and privacy-enhanced as well as LPDA is lightweight. Homomorphic Encryption (HE) can protect the data, but it is too slow especially multiplication. In addition, existing data mining studies using encrypted data without addressing this problem. B. K. Song et al [25] propose a data mining algorithm through logical gates. They design logic of atomic operations for encryption and apply logic to well-known algorithms.

4. Lists of items

This section provides the detailed discussion about the proposed PPDM framework.

4.1 System Overview

Data mining techniques are widely accepted for designing a number of personalized and secure applications which can be used for decision making and predictions. In these applications, a significant amount of data is used by learning algorithms [26]. But, at times the information is deficient and can affect the dynamic cycle. Accordingly, to finish the information, it is needed to assign the data from other genuine sources. Thusly, unique stack holders are club their information and dig it for viable dynamic. In this context, the parties (who combined their data) are worried about the confidentiality, security, and privacy, because, the data may contain private data, and discloser can impact on data owner's reputation. Therefore, to deal with such issues during the collaborative data mining and decision-making process, the PPDM is used. In this model, a data sanitization process is used to prevent security and privacy issues. But not all the methods offer higher data utility, less time and space requirements. Therefore we need some key improvements on existing PPDM models.



(A) (B)

Fig. 1. demonstrate two PPDM Architecture (A) shows the privacy management on server side (B) shows the privacy management at client end

4.2 Methodology

In order to conduct experiments, we have two different kinds of architectural options. The first architecture is demonstrated in figure 1(A). In this model both kinds of data accepted i.e. horizontal and vertical partitioned. Additionally, the model supports multiple data owners. Thus in top layer of model the data owners are demonstrated. The agreed parties submit their data on a central server.

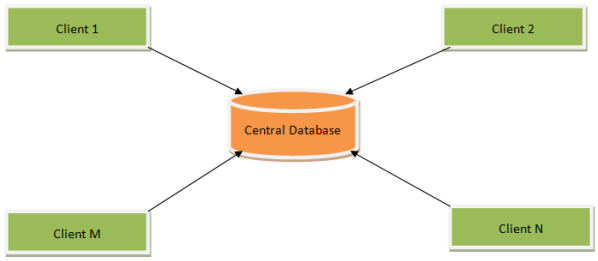


Fig. 2. Central Authority

In server the data is combined to produce a new dataset. Further noise is introduced on data to transform actual values. This process is call data sanitization. Due to data sanitization in a common place, the data manipulated uniformly. Further the data is used with data mining model. This model has some major draw backs such as privacy management among all the concerned parties. Additionally the private data recovery is a complex task. Thus by considering the server entity as semi-trusted a new model is proposed as described in figure 1(A). In this model it is assumed that the data analytic server is not completely trusted. Therefore, at the client end, before data submission the data is sanitized. Then data mining is carried out on one or more servers. The server includes two phases organizing the data, and processes the data. When the multiple servers are involved then we need to collect the data mining outcome and combine them for publishing.

4.3 Handling Vertical Partitioned Data

In PPDM multiple parties are involved but, no one has complete data. Additionally no party have equal amount of attributes. Therefore, parties are tried to combine own part of data for mining and decision making.

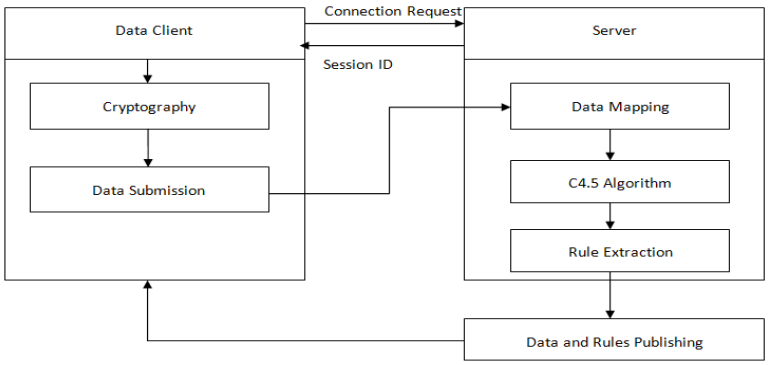


Fig. 3. PPDM Data Model for vertically partitioned data

To understand this scenario let us take an example, there is an institute with three departments A, B, and C. The authorities want to perform mining on data of students. Thus, all departments submitting their data to an authority and perform mining. Consider table 2 for more understanding [27].

Table 2 vertical partitioned data

Department A		Department B		Department C		
S1	S2	S3	S4	S5	S6	Class label

In this table S1, S2, ….., S6 are the subjects and their marks which are contributed by three departments to aggregate data in a common place. Figure 2 demonstrate how the N number of parties is connected to a server. That server can be a trusted or semi-trusted authority who is responsible to securely collect and process the entire data and mine the decision rules. The proposed data model for vertically partitioned data is demonstrated in figure 3. Additionally, the used components in this diagram are explained.

Data Client: the framework is a multiparty computation; therefore the data suppliers are the primary component of the system. Each client contributes their owned part of attributes with the semi-trusted authority. The authority is responsible to combine the data and apply the data mining technique.

Server: The server is the trusted authority who is responsible to process the entire data. Therefore, when a party wants to join the framework the server assign an ID to client. This ID is a random and unique number for each session. This ID is also used for encryption process during sensitization of data.

Cryptographic security: In the adopted PPDM model clients are sensitizing the data before submission. That feature offer the control on security and privacy to the end client. In order to sanitize data efficiently a cryptographic algorithm is proposed.
This model is combining the goodness of both AES and SHA1 algorithms. Figure 4 demonstrate the proposed cryptographic infrastructure. This cryptosystem accepts two parameters first server assigned session ID and data, which is needed to be sanitize. The session ID is used with the SHA1 algorithm to generate 160-bit alphanumeric hash codes. This hash code is processed using a key generation process, where the 160-bit hash is reduced to generate the 128-bit key. Further, AES algorithm accepts client data and SHA1 based cryptographic key and generates the cipher text. The steps of data encryption are demonstrated in table 3.

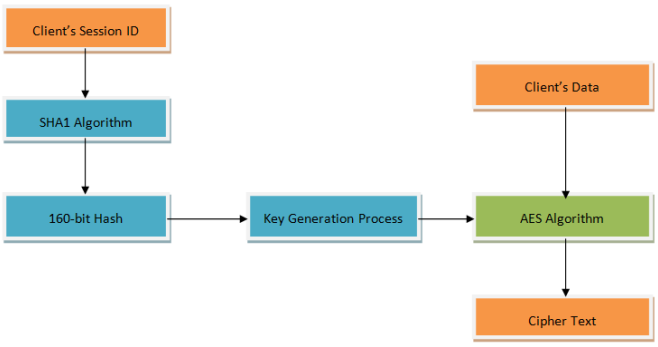


Fig. 4. Cryptographic technique

Data submission: the ciphered data is submitted over the server in terms of encrypted attributes and a readable class labels. The server organized all data into a common dataset, and a common class label.

Attribute mapping: however the ciphered data is not in human-readable format, therefore, the attribute values of the dataset is mapped into a symbols. The manipulated attributes of the dataset are used in the next processes.

Table 3 Cryptographic Algorithm

Input: User Session ID S, Data to encrypt D
Output: Cipher Text C
Process:

1. $k_{160} = SHA1.GenrateHash(S)$
2. $k_{128} = DiscardLSB(K_{160}, 32)$
3. $C = AES.Encrypt(D, k_{128})$
4. return C

C4.5 algorithm: in order to mine decision rules for decision making process by the combined dataset the C4.5 decision tree algorithm is used. The C4.5 algorithm is also known as J48 decision tree, which is an extension of the ID3 algorithm. The ID3 involves pruning steps to reduce the tree size and ambiguity to provide the higher accuracy. The C4.5 decision tree usage the learning samples to prepare decision tree after training, and after preparing the decision tree it usages to predict the similar pattern. The decision tree can be transformable into the "IF THEN ELSE" rules. In this work the C4.5 algorithm is applied. The C4.5 tree computes information gain (IG) for splitting data. Additionally to calculate the IG we need to measure entropy. The entropy of a dataset D with two class labels, i.e. True and False, can be defined as:

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

Where, P (T) is probability of True, and P (F) is probability of False.

In order to produce small size or to control the depth of tree we need to select of best features. These best features are also termed as attribute with minimum entropy and the IG is known as drop in entropy. That also helps to understand the relation among attributes. The IG, Gain (E, A) for attribute A is measured as,

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

The IG also decides the positions of attributes in the tree. Nodes with maximum gain, which are not considered yet, are used in tree development process. That is beneficial for the following purpose:

1. To generate a small-size tree.
2. To accomplish the preferred level of utility

Rule extraction: the C4.5 tree use dataset on server and prepares the tree. The developed decision tree is further transformed into the "IF THEN ELSE" rules. In tree the nodes demonstrate the attributes, edges describe the values of the attributes, and finally leaf nodes shows the decision. The decision tree branches are used for producing rules. To understand the process of decision rule generation let us consider an example. Let using a dataset C4.5 algorithm produces a decision tree as given in figure 5.

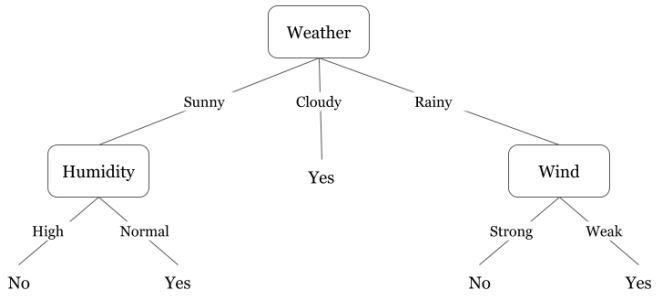


Fig. 5. Example decision tree

The tree contains the rules in their branches. Thus, by traversing the branches till the leaf node we can extract the "IF THEN ELSE" rules. Using the above tree we can prepare the following rules:

"IF 'Weather'= ='Sunny' and 'Humidity' = = 'Normal' THEN decision = 'YES'"

Rule publishing: the extracted rules at the server are needed to distribute among all the clients. Thus to distribute the rules, we perform a reverse mapping to create initially encrypted rules. After mapping of the rules in cryptographic format, are sent to all the parties. The clients capture the encrypted rules from server and use the decryption algorithm and session key. Using this process client only get the part of rule is recoverable only the submitted by the client.

Algorithm Steps: the model can be summarized using the algorithm steps as described in table 4. The algorithm accepts number of party data as input in vertically partitioned format. Additionally after data processing it returns sanitized data S and the data utility and performance. The server initiates a loop for all the parties, each party send a connection request to the server and establish connection then server sent a session ID (SID) to the connected party. When client receive the SID, then client generate the key for encryption K. Next, the client encrypts the data using AES algorithm and the key K. The encrypted data is generated by all clients thus S_i is the share of i^{th} client. Thus sanitized data$S = \{S_1, S_2, \dots, S_n\}$. During this process when the first party includes their data to server dataset then entire attributes are copied to sanitize dataset S, but the class attributes are included at last.

Table 4 Algorithm for Vertical partitioned data

Input: Number of Participant N, Partitioned Data $V_n = \{V_1, V_2, \dots, V_n\}$

Output: Sanitized Data S, Data Utility U

Process:

1. for$(i = 1; i < N; i++)$
    a. Client_Send_Connection_Request
    b. if$(connection == True)$
        i. $SID = Server.GenrateSessionID$
    c. End if
    d. $K = SHA1.GenrateKey(SID)$
    e. $S_i = AES.Encrypt(V_i, K)$
    f. if$(i == 1)$
        i. $S = S.Add(S_i)$
    g. else
        i. $S = S.Add(S_i, ClassAttribute)$
    h. End if
2. End for
3. $T_m = C45.Train(S)$
4. $U = T_m.evaluateModel$

5. Return U

Thus all the parties have the different set of attributes but a common class attribute. Now, the server mines the data using the C4.5 decision tree and generates the classification rules T_m. The T_m is collection of the decision tree rules. Finally the model is evaluated to obtain the performance or measure the utility of sanitized dataset.

4.4 Handling Horizontal Partitioned Data

As we know in most of the PPDM frameworks have multiple clients agreed to combine their data for analysis. In this environment all the parties has the similar attributes. This nature of data organization is known as the horizontal partitioning of data. In order to understand we can take an example where three different departments of educational institute are want to combine their data. Therefore to aggregate the data from all the departments, the horizontal partitioned architecture is used. Let the department A has table 5, which has the same data attributes as other two departments but the A has 80 instances, B has 120 instances and C has 100 instances with the similar attributes. Therefore, in horizontal partitioned system a total of 300 data instances. Here, we can also use the same architecture as previous. Therefore, the similar process is used in combining encrypted data, as well as their distribution.

Department A:

Student name	Age	Marks	Grade

Table 5 horizontal partitioned data for client A

Department B:

Student name	Age	Marks	Grade

Table 6 horizontal partitioned data for client B

Department C:

Student name	Age	Marks	Grade

Table 7 horizontal partitioned data for client C

Algorithm steps: the similar algorithm has been used for horizontal partitioned data mining. The difference among both the algorithm is that vertically partitioned data remove the class attribute form all the parties and club the data based on a common class label. But in horizontal partitioned data mining include the own class labels. The algorithm steps are described in table 8.

Table 8 Horizontal partitioned data Mining

Input: Number of Participant N, Partitioned Data $V_n = \{V_1, V_2, \dots, V_n\}$
Output: Sanitized Data S, Data Utility U
Process:

1. for(i = 1; i < N; i + +)
    a. Client_Send_Connection_Request
    b. if(connection == True)
        i. SID = Server. GenrateSessionID
    c. End if
    d. K = SHA1. GenrateKey(SID)
    e. S_i = AES. Encrypt(V_i, K)
    f. if(i == 1)
        i. S = S. Add(S_i)
    g. End if
2. End for
3. T_m = C45. Train(S)
4. U = T_m. evaluateModel
5. Return U

5. Results Analysis

The aim is to provide full control at the client to securing their own data. Thus two variants of PPDM model are proposed. The experimental evaluation of both the models is explained.

5.1 Experimental Scenario

The aim is to reduce the performance variation between the proposed PPDM models and the baseline model for optimal data utility after data sanitization process. The models are implemented and performance is evaluated to compare in two different experimental scenarios:

1. **Implementation of a cryptographic PPDM model:** in this experiment a model is proposed, which includes a cryptographic technique based on SHA1 and AES encryption. In addition, the model usages the C4.5 decision tree for mining the rules from data. The experimental analysis and comparison with base line model is provided.
2. **Implementation of a lightweight and accurate PPDM model for horizontal and vertically partitioned data:** in this approach we reduce the deviation of the previously model. Additionally, use both kinds of data partitions to enable more effective and secure data processing model. Thus, previous model is optimized for accepting both kinds of data partitions. Additionally provide the technique to reduce the attributes during submission to client, which reduces the resource consumption issue of the initial PPDM model. The comparative study among previously model, modified model and the base line model have been provided.

5.2 Evaluation

In order to improve the data utility and the performance in terms of time and memory utilization the PPDM model. The model is modified for developing LWE2EC. This section explains and compares the performance of the proposed LWE2EC model with base line and previously introduced model.

Accuracy of the LWE2EC model, Cryptographic model and the C4.5 is compared using table 9 and figure 6(A). The X axis includes the number of sample patterns and Y axis shows the

accuracy, which is measured in terms of percentage (%). Similarly the percentage error rate of the PPDM models is demonstrated in figure 6(B) and table 9. Here the Y axis shows the error rate in percentage. All the algorithms demonstrate the similar behavior in most of experiments for both the parameters. But in order to compare both the methods with baseline we also utilized mean accuracy and error rate of all the implemented models. The mean accuracy of the models is computed using:

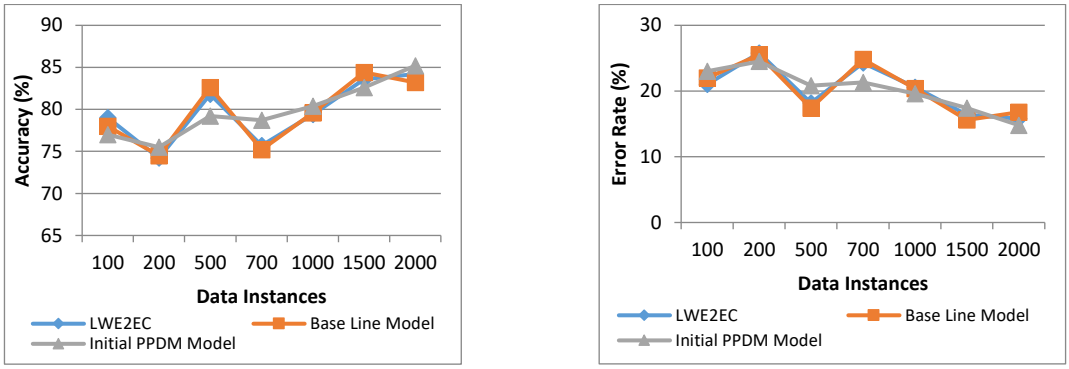
$$\text{Mean Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \text{Accuracy}_i$$

Where N is the number of experiments conducted and Accuracy_i is the accuracy of i^{th} experiment.

Additionally to measure the mean error rate the following equation were used.

$$\text{Mean Error Rate} = \frac{1}{N}\sum_{i=1}^{N} \text{Error Rate}_i$$

Where, N is the number of experiments and Error Rate_i shows the error rate of i^{th} experiment.



(A) (B)

Figure 6: demonstrate the comparative performance of two variants of PPDM (A) shows the Accuracy in (%) and (B) shows the error rate in (%)

According to the mean accuracy the initially PPDM technique shows higher accuracy as compared to LWE2EC and C4.5 decision tree algorithm. Because it is composed and sanitized on server by which the data include the noise uniformly. Similarly mean error rate is demonstrated in figure 7(C) and mean accuracy in 7(A). Here the X axis includes the different methods and Y axis includes the error rate and accuracy in terms of percentage (%). According to the performance of algorithms the initial PPDM model produces less error rate as compared to other two models. But the performance of C.45 algorithm and the LWE2EC model is closer. But we need a method that keeps the Data utility similar, thus we also computing the deviation of accuracy and error rate from the base line model. The deviation of the model is measured using:

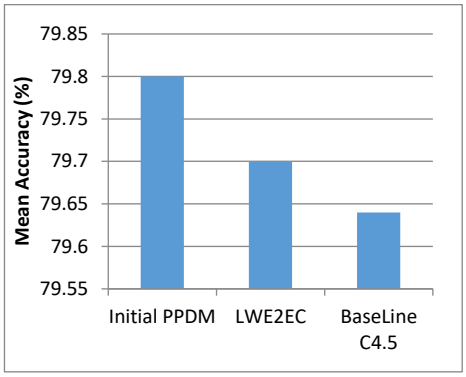
Table 9 Accuracy and Error Rate in Percentage (%)

Dataset	LWE2EC		Initial PPDM Model		Base Line Model	
Instances	Accuracy	Error Rate	Accuracy	Error Rate	Accuracy	Error Rate

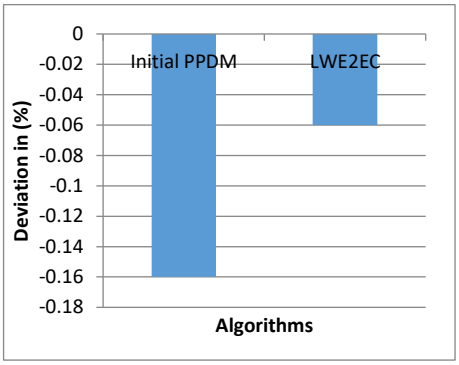
100	79	21	77	23	78	22
200	74.2	25.8	75.5	24.5	74.5	25.5
500	81.8	18.2	79.2	20.8	82.6	17.4
700	75.7	24.3	78.7	21.3	75.2	24.8
1000	79.4	20.6	80.4	19.6	79.6	20.4
1500	83.6	16.4	82.6	17.4	84.4	15.6
2000	84.2	15.8	85.2	14.8	83.2	16.8

$$\Delta = \text{BaseLine Model} - \text{actual performance}$$

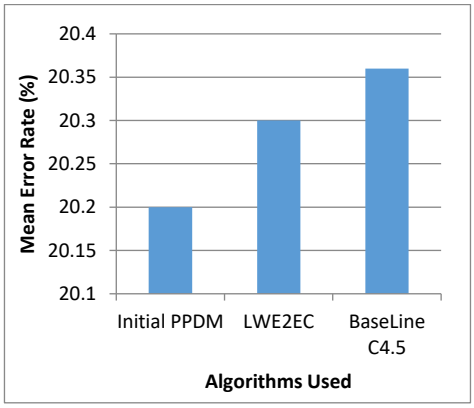
The difference between the proposed techniques and baseline model is demonstrated in figure 7(B) and 7(D) respectively. According to the results, LWE2EC model provides less fluctuation as compared to initial model. Suppose when we do nothing with the data and use with the C4.5 algorithm then the decision tree actually recognize 80% of samples.



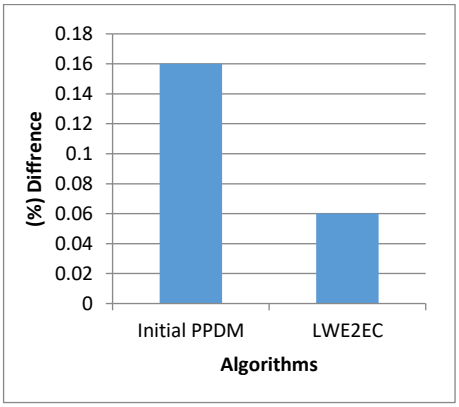
(A)



(B)



(C)



(D)

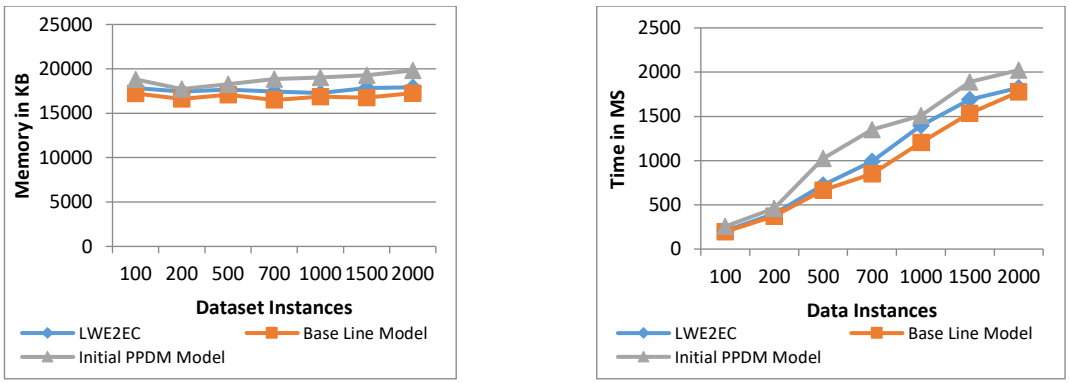
Figure 7 shows the mean performance of the PPDM models where (A) shows the mean accuracy (B) shows the deviation in accuracy (C) shows the Mean error rate and (D) shows the deviation in error rate

After data sanitization, the data is modified. Thus the utility of data is affected, therefore if a model manipulates the data and produce new dataset, then the same algorithm results 81% or 79% of correct recognition, so we can say some of the samples are change their behavior. The size of such data is 1%, but if the model results 85% or 75%, then we can say the model can change the behavior of actual data patterns more than 5%. The change in behavior of data can impact the performance of utility of data. So, we need the difference in performance is closer

to 0. Thus, we can say the LWE2EC model produce the manipulated data but not disturb the utility of data majorly. Thus the LWE2EC shows higher utility of data and achieving the higher privacy. In other terms the LWE2EC model provides much similar results as the baseline model. The memory usages of PPDM models are demonstrated using figure 8(A) and table 10. The Y axis contains the memory usages in terms of KB (kilobytes). According to the obtained results the LWE2EC model consumes higher amount of main memory as compared to C4.5 algorithm. Basically the encryption needs additional memory to encrypt and decrypt data thus the main memory usages is increases. Thus in order to know the mean performance of all the PPDM models, the mean memory usage is given in figure 9(A). In order to calculate the mean memory usages the following equation were used:

$$\text{mean memory} = \frac{1}{N}\sum_{i=1}^{N} \text{memory}_i$$

Here, N is the number of experiments and memory_i is the memory used in i^{th} experiment. According to the results the cryptographic model needs higher degree of memory and the LWE2EC model shows less usages.



(A) (B)

Figure 8 shows the performance of PPDM model where (A) shows the memory usage in terms of kilobytes and (B) shows the Time consumed in milliseconds

Thus the performance variation are also measured and reported in figure 9(B). That demonstrates the difference between the actual and after privacy preserving. According to the results the LWE2EC model consumes less memory as compared to initial PPDM model. The difference between the baseline model and LWE2EC is near about 0 but the difference between initial PPDM and base line model is significant. Thus the LWE2EC is acceptable due to less difference. Next the time requirements of both the models are demonstrated in figure 8(B) and table 10. The Y axis shows the Time utilized in terms of MS (milliseconds). According to the results the base line model C4.5 demonstrates less time utilization. In order to find the difference mean time utilization were also measured. The mean time requirements are calculated using:

$$\text{Mean Time} = \frac{1}{N}\sum_{i=1}^{N} \text{Time}_i$$

Table 10: Performance in terms of memory and time consumed

Dataset	LWE2EC		Initial PPDM Model		Base Line Model	
Instances	Memory	Time	Memory	Time	Memory	Time
100	17829	206	18829	256	17232	195
200	17429	398	17729	459	16624	372
500	17664	725	18264	1025	17074	665
700	17436	991	18836	1349	16488	850
1000	17279	1396	19027	1509	16864	1204
1500	17824	1689	19282	1889	16754	1534
2000	17922	1823	19822	2023	17268	1776

Where N is the number of experiments and Time_i is the time consumed during i^{th} experiment.

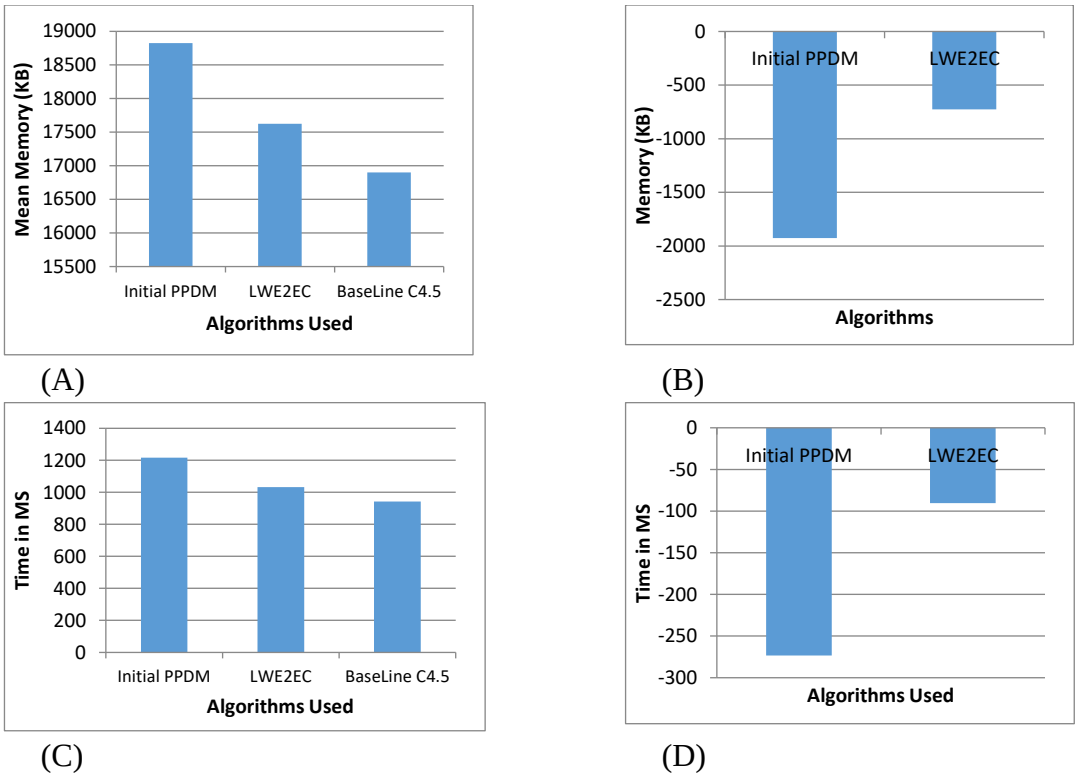


Figure 9 shows the performance of models in terms of (A) mean memory consumed (B) variation between base line and proposed PPDM model in memory usages (C) demonstrate the mean time consumption and (D) shows the variation in time

Mean time requirements are explained in figure 9(C), here the Y axis shows the time consumed. According to the results the initial PPDM consumes higher amount of time. Additionally, LWE2EC reduces the amount of time requirements. The variation with respect to the C4.5 algorithm is also measured and reported in figure 9(D). That is the total difference of time

which is additionally consumed from the baseline algorithm. We found the time variation of the LWE2EC model demonstrate the efficient outcome as compared to initial PPDM model.

6 Conclusion and Future Work

This section demonstrates the experimental and observational facts which are noticed during design and evaluation of the system. In addition the future extension is also proposed.

6.1 Conclusion

The data is become more expensive then gold, therefore in online and business applications security and privacy is the main concern of the data owners. However, in some applications organizations need to combine the data with the third party to get business insights as well as planning and critical decision making. Therefore, we need to involve the third party service provider to help in mining and decision making process.

In this scenario the PPDM is the best and acceptable solution for mining data with security and privacy. But the structure of data for collaboration is a great challenge, because there are a fewer models which can consider horizontal and vertical partitioned data. Therefore the paper includes a study of the PPDM models for both the kinds of data structures. Then a cryptographic model is proposed to secure data at the network level, and third party. Additionally, multi-party vertically partitioned data is used to evaluate the working of the model with respect to the baseline model. Further, to deal with horizontal and vertically partitioned data both the cryptographic model is modified. Both the models are implemented and compared against the baseline modes for investigation of the data utility. The experimental results are summarized in table 11.

Table 11 Performance summary

S. No.	Parameters	Base line	Vertical PPDM	LWE2EC
1	Accuracy	Ref	Fluctuating	Close to ref
2	Error rate	Ref	Higher	Close to ref
3	Memory usages	Ref	Higher	Close to ref
4	Time consumed	Ref	Higher	Close to ref

According to experimental results the proposed methods are work well for cloud based systems. Additionally using fewer modifications we can achieve a common framework for mining both kinds of data vertically as well as horizontal partitioned data.

6.2 Future Extension

The paper is accomplishing a secure and privacy preserving technique and exploration of new opportunities in security and privacy for designing new generation data centric applications. Thus the following future proposal is offered.

1. The current work involve the cryptographic scenario, in near future we can extend the model using new security technique
2. The model can also be extended to offer a data driven service where anyone can collaborate their data and can get the relevant insights

3. Apply the model in a real world applications such as IoT enabled applications to mine decisions and automate the working of devices

References

1. N. Kashyap, Dr. V. Bhattacharjee, Security in Privacy Preserving Data Mining, Inter. Jour. Of Engg & Comp. Sci., ISSN:2319-7242, Volume 4 Issue, Page No. 11698-11703, (2015)
2. R. Natarajan, Dr. R. Sugumar, M. Mahendran, K. Anbazhagan, A survey on Privacy Preserving Data Mining, Inte. Jour. of Adv. Rese. in Comp. & Comm. Engg, Vol. 1, Issue 1, (2012)
3. R. Mendes and J. P. Vilela, Privacy-Preserving Data Mining: Methods, Metrics, and Applications, Volume 5, 2017, 2169-3536 2017 IEEE
4. L. Xu, C. Jiang, J. Wang, J. Yuan, Y. Ren, Information Security in Big Data: Privacy and Data Mining, VOL. 2, 2014 2169-3536, IEEE
5. P. Jain, M. Gyanchandani, N. Khare, Big data privacy: a technological perspective and review, J Big Data (2016) 3:25 DOI 10.1186/s40537-016-0059-y
6. Y. A. A. S. Aldeen, M. Salleh, M. A. Razzaque, A comprehensive review on privacy preserving data mining, Springer Plus (2015) 4:694 DOI 10.1186/s40064-015-1481-x
7. V. S. Susan, T. Christopher, Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes, Springer Plus (2016) 5:964 DOI 10.1186/s40064-016-2490-0
8. Mrs. S. Shelke, Prof. B. Bhagat, Techniques for Privacy Preservation in Data Mining, Inter. Jour. of Engg. Rese. & Tech., Vol. 4 Issue 10, October-2015
9. M. Arafati, G. G. Dagher, B. C. M. F. Sis, P. C. K. Hung, D-Mash: A Framework for Privacy-Preserving Data-as-a-Service Mashups, 2014 IEEE 7th Inter. Conf. on Clo. Comp. (CLOUD)
10. L. Li, R. Lu, K. K. R. Choo, A. Datta, J. Shao, Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases, 1556-6013 (c) 2016 IEEE
11. C. W. Lin, T. P. Hong, H. C. Hsu, Reducing Side Effects of Hiding Sensitive Item sets in Privacy Preserving Data Mining, Hind. Publ. Corp., e Scie. Wo. Jour. Vol. 2014, Art. ID 235837, 12 pages.
12. X. Shu, D. Yao, E. Bertino, Privacy-Preserving Detection of Sensitive Data Exposure, IEEE Trans. on Infor. Fore. & Sec., VOL. 10, NO. 5, MAY 2015
13. J. Li, Z. Liu, X. Chen, F. Xhafa, X. Tan, D. S. Wong, L-Enc DB: A lightweight framework for privacy-preserving data queries in cloud computing, Know.-Bas. Syst. (2014)
14. Q. Zhang, L. T. Yang, Z. Chen, Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning, IEEE Tran. on Comp., Vol. 65, No. 5, May 2016
15. X. Yi, F. Y. Rao, E. Bertino, A. Bouguettaya, Privacy-Preserving Association Rule Mining in Cloud Computing, ASIA CCS'15, April 14–17, Singapore. ACM 978-1-4503-3245-3/15/04
16. K. Xu, H. Yue, L. Guo, Y. Guo, Y. Fang, Privacy-preserving Machine Learning Algorithms for Big Data Systems, 2015 IEEE 35th Inter. Con. on Distr. Comp. Sys.
17. Z. Fu, F. Huang, K. Ren, J. Weng, C. Wang, Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data, IEEE Trans. on Info. For. & Sec., VOL. 12, NO. 8, AUG. 2017

18.J. Oksanen, C. Bergman, J. Sainio, J. Westerholm, Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data, Jou. of Tran. Geo. 48 (2015) 135–144
19.T. Y. Wu, J. C. W. Lin, Y. Zhang, C. H. Chen, A Grid-Based Swarm Intelligence Algorithm for Privacy-Preserving Data Mining, Appl. Sci. 2019, 9, 774; doi:10.3390/app9040774
20.Y. Li, W. Xu, PrivPy: General and Scalable Privacy-Preserving Data Mining, KDD ’19, August 4–8, Anchorage, AK, USA ACM ISBN 978-1-4503-6201-6/19/08
21.J. C. W. Lin, P. F. Viger, L. Wu, W. Gan, Y. Djenouri, J. Zhang, PPSF: An Open-Source Privacy-Preserving and Security Mining Framework, 2018 IEEE Inter. Conf. on Da. Min. Wor., DOI 10.1109/ICDMW.2018.00208
22.M. L. Merani, D. Croce, I. Tinnirello, Rings for Privacy: an Architecture for Large Scale Privacy-Preserving Data Mining, IEEE Tran. on Par. & Dist. Syst., 1045-9219 (c) 2020 IEEE
23.M. M. Siraj, N. A. Rahmat, M. M. Din, A Survey on Privacy Preserving Data Mining Approaches and Techniques, ICSCA '19, February 19–21, Penang, Malaysia, ACM
24.R. Lu, K. Heung, A. H. Lashkari, A. A. Ghorbani, A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT, 2169-3536, 2017 IEEE, VOL. 5
25.B. K. Song, J. S. Yoo, M. Hong, J. W. Yoon, A Bitwise Design and Implementation for Privacy-Preserving Data Mining: From Atomic Operations to Advanced Algorithms, Hind. Sec. & Comm. Netw. Vol. 2019, Article ID 3648671, 14 pages
26.R. K. Dwivedi, R. P. Bajpai, Use of Data Mining in the field of Library and Information Science : An Overview, 2nd Inter. CALIBER-2004, New Delhi, 11-13 February
27.K. Ahuja, N. Sharma, D. K. Mishra, R. K. Vyas, Investigation of Privacy-Preserving Data Models and Contributions, Proc. of the 13th INDIACom-2019; IEEE Conference ID: 46181.