Privacy Preserving Data Mining: A Comprehensive Survey

Ritika Lohiya
Assistant Professor
Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat

Ankita Mandowara
Assistant Professor
Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat

Rushabh Raolji
Assistant Professor
Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat

ABSTRACT
Privacy preserving data mining has emerged due to large usage of data in organizations for extracting knowledge from data[1]. Big data uses centralized as well as distributed data and mines knowledge. Privacy preservation of data has become critical asset due to malicious users and society issues. It is very crucial nowadays to maintain balance between ensuring privacy and extracting knowledge. These areas is burning domain for researchers till now because no such research has been done that out performs all the techniques in privacy preserving data mining. Privacy preservation is classified into many categories like data modification, data distribution, data hiding and data encryption. For performance measuring, evaluation criteria like information loss, computational overhead, data utility etc are considered. Data modification techniques mainly focus on adding errors to data or results into output which degrades the accuracy of data mining algorithm. In case of critical analysis of data, crypto graphical approaches in privacy preserving data mining which has no loss of information but overhead of computation and communication have been adopted. PPDM includes homomorphic encryption, Shamir's secret sharing scheme, oblivious transfer and many other cryptography techniques. Challenges in this area include, higher computational and communication cost. At last, most advanced, functional encryption concept in privacy preservation have been included. Functional encryption provides higher level of security as well as privacy to data. It only allows learning output of function without revealing anything else.

Keywords
FE, PPDM, STTP, TTP

1. INTRODUCTION
In today's information age, data collection is ubiquitous, and every transaction is recorded somewhere. The resulting data sets can consist of terabytes or even petabytes of data, so, efficiency and scalability are the primary consideration of most data mining algorithms [2].

Naturally, ever-increasing data collection, along with the goal of data mining, i.e. to extract knowledge from data leads to privacy concerns. User friendliness of data mining results lead us to protect against leakage of individual's private information. For example, with the help of join operation on databases which are private and publicly avail- able, private information about any citizen can be leaked.

Data can be homogeneous, or, it can be heterogeneous in nature. In distributed environment, where risk is higher in processing data for mining, as it has more number of issues like secure line for communication, honest parties involved, third party behavior etc. and so higher level of security is needed to overcome such issues. Due to higher level of security and privacy provided by cryptography based approaches, these are applied to provide privacy as well as security. In privacy preserving distributed data mining, two types of communication models are used, which are, Trusted third party and Collaborative Processing[17]. In case of trusted third party, all the computation and key distribution party is handled by central authority while in later model; parties themselves take care of aggregation of results. Later approach exhibits higher cost due to higher computation. Advanced development of cryptography, like, Functional Encryption[43] has proven significantly helpful in data privacy. Though, it is based on many assumptions of cryptography, it is highly secure and efficient scheme till now in cryptography. This paper has analyzed the concept of Functional Encryption like Attribute based encryption, Identity based encryption and Inner product based encryption[46]. Functional Encryption allows receiver to learn the output of the function defined, instead of data which is encrypted.

This paper is divided into 8 sections. Section 2 explains how ppdm techniques are classified based on the operations on data or mining algorithm. Section 3 illustrate the research done in the field of ppdm. Section 4 explains, techniques in ppdm and its research areas. Section 5 illustrates the research done in the field of privacy preserving distributed data mining. This paper also surveys new concept which provides data privacy, Functional Encryption(FE). Evaluation criteria for privacy preserving data mining algorithm are explained in section 7. Section 8 contains conclusion and future work in this area.

2. APPROACHES FOR CLASSIFICATI-ON OF TECHNIQUES
There are many approaches which have been adopted for privacy preserving data mining. K and B Srinivasan Rao have explained the dimension of privacy preserving data mining. Their classification has been done based on the following dimensions [3]:

- data distribution
- data modification
- data mining algorithm
- data or rule hiding
- privacy preservation

The primary measurement alludes to the distribution of information. A percentage of the methodologies have been produced for incorporated information, while, others allude to

conveyed information situation. Appropriated information situations can likewise be delegated level information, circulation and vertical information appropriation. Level conveyance alludes to these situations where diverse database records live in better places, while vertical information dissemination, alludes to the situations where all the qualities for distinctive traits dwell in better places. Second dimension deals with the modification of data before producing the results. Modification includes three measurements:

Perturbation, includes adding error to the data or changing the value of data (i.e. change 1s into 0s and vice versa). Two possible perturbation are additive and multiplicative Sampling, includes taking samples from data and producing the results. The third measurement alludes to the information mining calculation, for which the information change is occurring. This is really something that is not known in advance, but rather it encourages the investigation and configuration of the information concealing calculation. The issue of concealing information have been incorporated for a blend of information mining calculations, in the future exploration plan[3]. The fourth dimension refers to hiding the sensitive attributes or sensitive rules which lead to the problem of data inference. And the final dimension refers to the technique for protecting privacy of data which includes heuristic technique, cryptography technique and condensation approach. It is very important to realize that data modification leads to result degradation and data encryption requires more computational time. So the balance between these techniques are required to get the optimum results.

3. LITERATURE STUDY

In the field of cryptography, researchers have given many fruitful results in last decades. But due to increasing demand of privacy for mining results, researchers are till now trying to give their best solutions to the issues. Many firm solutions have been seen in this field which are mentioned below in the survey. In privacy preserving distributed data mining, knowledge discovery is done by processing the data resided in remote sites[5]. Every database struggles against privacy and security scenarios. Privacy preserving distribute data mining provides both privacy and security to knowledge discovery process.

Privacy preserving data mining consists of two approaches, the modification based and cryptography based approaches. Later provides higher level of privacy as well as security in terms of distributed scenario. Former approach is useful in scenarios where data is at one place and only modification of results or data ensure the privacy factor. Later approach will be focused as it provides both security and privacy but do not have any standardize framework and also it has higher computational and communicational overheads. Cryptography based approaches has many flavors, but here following techniques have been focused which are suitable in this case. Homomorphic encryption based, oblivious transfer, elliptic curve cryptography, secret sharing based technique are to ensure the secrecy and privacy of the database. Also, oblivious transfer has higher communicational and computational overheads and therefore they are not suitable for larger databases. So focus will be on the remaining techniques to be applied in the research. These techniques have been used along with distributed data mining algorithms to achieve best results for the research. Homomorphic encryption has two approaches, symmetric encryption and asymmetric encryption. Later is very useful in distributed scenarios as it has separate mechanisms if the third party exists, former approach has lower overheads but sometimes is fails to provide security factor. Secret sharing has approaches like Shamir's secret sharing scheme and Verifiable secret sharing scheme. Both have their advantages and disadvantages which need to be worked upon.

Recent research in privacy preserving data mining is burning topic due to remotely located database, privacy laws and friendliness of data mining results. Privacy preserving data mining has many approaches which need to be explored to achieve qualitative and innovative results. At this age of big data, privacy concern has grown more for individuals as well as for organizations. When it comes to providing privacy and security, security increases cost which leads to degradation of overall scenario. The objective is to achieve qualitative research results in distributed environment by applying data mining algorithm with provision of security. Efforts will be put to reduce the cost of security. Adopted approaches are namely, homomorphic encryption, secret sharing schemes, oblivious transfer. In earlier stage, all the techniques of security provision will be analyzed with data mining algorithm. Flaws and limitations will be found out. With that information, framework of privacy preserving distributed data mining will be produced. At the end, the proposed framework will be implemented and results will be produced.

Table 1. Heuristic Based Approach

Author	Year	Approach	Findings
Samarati P.[19]	2001	First annonymization	Usage of Generalization and suppression. Provides identity disclosure.
Olivieria Zalane[10]	2002	Usage of sanitizing Data	Only applicable to association rule mining.
Zhong S. et.al.[5]	2005	Heuristic and cryptography based solution	Provides end to end security.

4. TECHNIQUES FOR PRIVACY PRESERVING DATA MINING

4.1 Anonymization technique

In the most basic form of PPDM, following is the table format given by data owner D(Explicit Identifier, Quasi Iden- tifier, Sensitive Attributes, Non-Sensitive Attributes) Explicit Identifier is a set of characteristics, for example, name and standardized savings number, containing data that expressly distinguishes record managers. Quasi Identifier is a set of qualities that could conceivably distinguish record holders. Sensitive Attributes comprises of delicate individual particular data, for example, sickness, pay, and handicap status and Non-Sensitive Attributes contain all the properties that do not fall into the past three classes. The four sets of properties are disjoint restricted. Anonymization[2] refers to the PPDM approach that seek to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed.

Table 2. Reconstruction Based Approach

Author	Year	Approach	Findings
Srikant and Agrawal[34]	2001	First technique proposed	Usage with classification approach
Dutta H. et. Al.[27]	2003	Data distortion using noise	Smaller noise gives good results.
Kamrakar and Bhattacharya	2009	Randomization and perturbation to modify data	Works good with centralized data. Biased towards privacy at the cost of data utility.
Xiaolin and Honglin[12]	2010	Amplifying matrix condition used	Better trade between usability and privacy
Vaidya and Clifton[18]	2003	Vertically partitioned data,clustering algorithm	Tradeoff between comp. cost and privacy
Bing Yang, Hi-roshi, nakagawa, jun sukama[26]	2010	Secure summation, homomorphic encryption	Compared to vaidyas protocol Communicati-on cost is low
Sankita and d c jinwala[16]	2013	Shamir's secret sharing scheme, collaborative model	Reduced cost of computation

4.2 Data Perturbation

This technique is classified into two categories 1) proba- bility distribution sampling 2) fixed data perturbation. The probability distribution type considers the database to be an example from a given data that has a given probability distribution. In given situation, security controlling technique replaces the first information by an alternate, example ,from the same distribution or by the appropriation itself. In the fixed data perturbation type, the estimations of the characteristics in the database, which are to be utilized for processing in- sights, are perturbed once and for all. The altered information perturbation strategies have been created solely for either numerical information or absolute information[3].

Following are the Data perturbation methods:

Noise additive perturbation, generally adds additive noises to attribute values. The data owner may not need all the column to be private so column based perturbation is applied. This technique is simplest form of perturbation which includes additive perturbation and noise additive perturbation.

Condensation-based perturbation aims at preserving co-variance matrix for multiple columns. *Random projection based* perturbation, refers to the technique of projecting a set of data points from the original multidimensional space to another randomly chosen space.

Table 3. Cryptographic Based Approach

Author	Year	Approach	Findings
Lindell and Pinkas[35]	2000	Secure multiparty computation	On ID3 algorithm. Higher computational cost

4.3 Cryptography Based Technique

Investigation in protection saving information mining began after 2000, but, foundation of cryptographic based technique was laid by Yao 1982[8]. In Yao's moguls issue, two tycoons need to figure out who is wealthier, yet without uncovering their wealth to the one another. Actually, the capacity to look at numerical information is essential in most information mining assignments. Yao's work launched research in secure multi- party processing which is the investigation of the class of capacities where two or more players can safely process on their joint inputs. This is carried out in a manner that only the last consequence of the reckoning will be uncovered to the gatherings. Specifically, no gathering will know the inputs of alternate gatherings. Mainly privacy preserving distributed data mining uses cryptography techniques to gather local models from local sites and aggregate into central model. This technique helps in accuracy of results but degrades the performance of data mining algorithm as the computational overhead is higher due to complexity of cryptography algorithms. Shamir's secret sharing[7], homomorphy encryption[11], secure multi party protocols[12], public key cryptosystems etc. algorithms are there in privacy preserving data mining.

Public key encryption plans are focused around higher computational expense and they oblige methods, for example, modular exponentiation of large numbers (in the order of 1k bits). on the contrary, it is extremely productive to figure secret shares when utilizing e.g. Shamir's secret sharing or the simple additive secret sharing. Public key encryption system creates cipher texts of at least 1024 bits. If the homomorphic properties of an encryption scheme are used then each input has to be encrypted in its own cipher text. Following table shows the research in this field.

Goldwasser-Micali, 1984 Because of its restrictive message development during encryption (i.e. every 1 bit plain- text is encrypt as ciphertext of every 1024 bits) Benaloh cryptosystem permits the encryption of bigger piece sizes at once Paillier cryptosystem encodes 1024-bit messages in ciper texts of no less than 2048 bits, which is feasible on the off chance include huge plaintexts. Secret sharing was introduced independently by Shamir[7]. Many other secret sharing schemes like Diffie hallman was also available in cryptography field but itallows to share secret key between

two parties with a prior establishment. Security is dependent on security parameter used in scheme. In Shamir's secret sharing schemes more number of parties can be involved and threshold t is decided to construct secret key. Less then t parties trying to reconstruct secret key will not be possible and they will fail to learn anything. In the multi-party scenario[20], there are protocols that enable the parties to compute any joint function of their inputs without revealing any other information about the data which is given as input. That is, applying the function while attaining the same privacy as in the ideal model.

5. PRIVACY PRESERVATION IN DISTRIBUTED DATA MINING

Distributed data mining permits distributed sites owning individual datasets to perform mining by joining their data. Data is scattered to different sites. By applying mining algorithm locally, Local result model is prepared. Results are then aggregated using either trusted third party or secret sharing schemes. In distributed scenario, many problem arises while processing data for mining, like different naming conventions for attributes at different sites, arbitrarily partitioned data across sites. There are many real world problems in distributed databases where privacy of data is major concern. First, different hospitals want to mine their databases for research purpose, putting privacy of their patient in danger. Second, different intelligent agencies want to mine their data without revealing any information about agencies or their operations. Due to these problems, different organizations cannot directly share or pool their databases without preserving privacy and ppdm aims to achieve this. For further understanding, following table shows the structure of algorithms used in privacy preserving distribute data mining.

5.1 Data Distribution Model

Firstly, discover how the data are partitioned while applying PPDM algorithm. The relational databases are the most commonly used databases in distributed scenario. Therefore focus is on different data partitioned model in context of the relational model. In horizontal partitioned[24] dataset, different site collects same types of columns about the different databases. For example two organization collects same type of database. However customer database for each database schema might be different. This database structure usually occurs in same organization or across similar do-mains. For example two medical institutes viz. PS medical college and LJ hospital, each of which collects information of their patients. attributes like patientcode, gender, occupation, age and disease are stored in both datasets. Merging two datasets gives more accurate predictive models. In vertically partitioned data, all organizations have the same objects, but different types of attribute. For example, database might be of following type: in some city big bazaar took their customers information buying tomatoes and potatoes. In the same city, d mart(competitor of Big bazaar) gathers information of customers buying beaf, onions. Now these two datasets have some linking information which helps in joining them. Mining of these two datasets outputs the buying behavior of customers in that particular city. For vertically apportioned database, it is expected that no organizations offer variables. With a specific end goal to match up a vertically partitioned[24] database, all organizations must have a worldwide identifier, for example, security number by government. Both the models are themselves useful in different scenarios. In some cases, arbitrarily portioned data sets are used which consist of very complex structure to handle.

Table 4. Privacy Preserving Distributed Data Mining(PPDDM)

PPDDM algorithm	
Data distribution	Horizontally, Vertically
Data Mining Algorithm	ARM, Classification , Clustering
Communication Model	Semi trusted third party, Collaborative Computation based
Cryptography based technique	Oblivious transfer, homomorphic encryption , secret sharing based

5.2 Data Communication Model

Here, third classification dimension, secure communication model, which generally refers to the interactive relation of the participants joining in the cooperative computation and the roles they play in the whole process of privacy preserving distributed data mining tasks. Another similar term is "coopetative" model [36]. This term stems from the word "coopetation", which was originally employed in social-economics to describe the situation that competing entities producing the same line of products and services have to cooperate with each other to improve the overall value of their market by means of making decisions based on the joint analysis of their private data. Similarly, distributed data mining tasks commonly feature a scenario where all the data holders participating in the joint computation on their individually private data sets naturally have the desire and interest to obtain the final result of the application. As the proprietary owner of their individual data set, it is understandable that each data holder is reluctant to share private information with other data holders. However, in order to reach the final result of the distributed data mining, they are ready and motivated to provide inputs to the computation, as long as the privacy requirements are met. Generally, most practical approaches to solve this scenario is to conduct the secure computation at one or more of the participants or at one or more third parties with the assumption that all of participants are semi-honest [39] and the third parties are semi trusted participants [39]. Here in, consider give the informal definition of both semi honest and semi-trusted. In [44], a semi-honest party (i.e. honest but curious) follows the rules of the protocol using its correct input, but is free to learn from what it sees during the execution of the protocol to compromise security. In [56], a third party is semi-trusted if it fulfills the following condition: the third party is trusted to provide some commodities or compute intermediate outcome of the computation based on encrypted input it receives; it follows the execution of the protocol correctly, just like all the other users as well, although it tries to learn and deduce some information from its own input and output.

Under such scenario and assumption, privacy-preserving distributed data mining problems can be solved mainly based on two types of secure computation model: One is based on Semi-trusted Third Party (STTP) model. Theoretically, the general secure multi-party computation protocols can be used to deal with any collaborative data mining problems, yet this kind of solutions are too inefficient when the database is huge in amount and the number of participants is large, due to its intricate and complicated design. On the other hand of the

spectrum, the trusted third party (TTP) model is too naive and straightforward, so that the privacy is compromised to a larger extent at the point of the TTP. Therefore, more practical solutions have been put forward in the past few years with respect to how to solve the privacy issues of distributed data mining more efficiently and accurately. Among them, two broad 35 streams of ideas are manifesting themselves: one is to introduce a semi-trusted third party, as compared to the trusted third party (TTP). In real world, it is much more feasible to find such a semi-trusted third party than to find a trusted third party. This semi-trusted third party can be implemented by means a miner, a mixer, or a commodity server, that all act in a semi-trusted manner. The other stream is based on Specific Secure Multi-party Computation under Semihonest assumption (SSMC). It aims at accomplishing efficient and accurate solution for the PPDDM problems. Under the semi-honest assumption, specific secure multiparty computation protocols are employed to deal with functions commonly used in data mining applications rather than the general secure multi-party computation protocol. These techniques include secure sum, secure set union, secure intersection, secure scalar product, etc. The advantage of such kind of protocols and tools lies in that they are designed to specially fit in with the data mining tasks, instead of any general functions. As the function for secure computation can be identified, the computing complexity is reduced greatly and a linear proportional cost can be obtained.

5.3 Oblivious Transfer

In cryptography, a oblivious transfer protocol (regularly condensed OT) is a kind of protocol in which a sender exchanges one of the possibly numerous bits of data to a collector, however stays secret regarding what piece (if any) has been exchanged. The first type of oblivious transfer[36] was presented in 1981 by Michael O. Rabin. In this structure, the sender makes an impression on the recipient with probability 1/2, while the sender stays oblivious in the matter of whether the collector got the message. Rabins oblivious transfer scheme is focused around the RSA cryptosystem. A more valuable type of oblivious transfer called 1-2 oblivious transfer was produced later by Shimon Even, Oded Goldreic and Abraham Lempel [2], so as to construct conventions for secure multiparty processing. It is summed up to 1 out of n oblivious transfer where the client gets precisely one database component without the server getting to know which component was questioned, and without the client knowing anything about alternate components that were not recovered.

5.4 Homomorphic Encryption

Homomorphic encryption system allows specific type of operations on encrypted data and outputs the results on encrypted data. If one wants to carry out operations like addition and multiplication on statistical encrypted data, there exist public key cryptography system for certain operations. The result is also an encrypted data. There are several partial homomorphic encryption system existing, which are less secure. Fully homomorphic encryption systems are more secure and provide secure computation on encrypted data. In recent development, in the field of cryptography, more secure encryption systems are developed for processing encrypted data and producing results but they have higher computation cost due to complexity of algorithms.

5.5 Sharing Based

Secret sharing or secret splitting amongst N number of parties proposed by shamir's secret sharing schemes allow N number of parties to share secret without any prior establishment of keys. If secret threshold is kept t, then minimum t number of parties have to gather to compute or to know about the whole secret. Shamir's secret sharing system shares secret using polynomials and random values with arbitrarily selected bits and then construct it back using lagranges interpolation equation.

Table 5. Survey Table

Elements	Complexity	Bytes exchanged	Communi-cation round	Scalability
Secure Communication Model				
STTP	Low	High	Low	High
SSMC	High	Low	High	Low
Data Mining Tasks:				
Classification	Low	N/A	N/A	High
Association Rule	Low	N/A	N/A	High
Clustering	High	N/A	N/A	Low
Privacy Preserving Technique:				
Homomorp-hic Encryption	Low	Medium	N/A	High
Oblivious Transfer	High	High	N/A	Low
Secret Sharing	Medium	Low	N/A	High
Randomizat-ion	Low	Low	N/A	High

6. NEW STUDY IN PPDM(FUNCTIONAL ENCRYPTION)

The paper has incorporated new study in the field of cryptography called functional encryption. It generates restricted keys to learn specific output of function on encrypted data, but learn nothing about the data. Encryption of data usually deals with securely sharing data over non secure network or data storage. Earlier in cryptography, if two parties wanted to communicate over non-secure channel, they had to do a prior establishment to encrypt their data for non-secure line. While this is acceptable for two party case but large number parties generate larger cost of communication and computation. Nearly thirty years ago, Diffie and Hellman gave solution to share secret key without any prior establishment or sharing any secret. Now it is time to study advanced topic in cryptography named Functional Encryption. In functional encryption[46] system, a decryption key allows user to learn a function of encrypted data. More precisely, in functional encryption system for function F (. ; .), a Third party holding MSK generates subkeys k that allow to compute function F (. ; .) on data which is encrypted. Dan Boneh and Amit Sahai gave a brief definition and security proofs of functional encryption system. They gave simulated definition of FE. Same authors have also defined data privacy and function privacy in functional encryption.

Definition 1. Functional encryption[?](FE) is defined for function F over (K, X) is a tuple of four PPT(probabilis- tic polynomial time) algorithm in the following manner: (setup, keygen, encr, decr) satisfying the following:

- SETUP(SP) is a p.p.t algorithm that takes input as security parameter and outputs master public key and master secret key (MPK, MSK).
- KEYGEN(MSK, C) is a p.p.t algorithm that takes input master secret key MSK and circuit and Outputs corresponding secret key Sk.
- ENCRYPT(PK, x) is p.p.t algorithm that takes input as master public key PK and an input message x and outputs ciphertext CT.
- DECRYPT(SK, CT) is deterministic algorithm that takes input as secret key SK, CT and outputs C(x).

6.1 Schemes of Functional Encryption

For applicability of functional encryption following two classes are defined

- Predicate encryption with index.
- Predicate encryption without index.

6.2 Predicate Encryption with Index

The study with simplest encryption case of Identity based encryption[46] in functional encryption and move towards advance path with attribute based encryption.

1.) *Identity based encryption:* In Identity based encryption cipher text and secret keys are connected with characters and a secret key can decrypt a cipher text if two of them are equivalent. IBE shows the first utility which isn't feasible from asymmetric key encryption. Boneh and Franklin and Cocks build first practical development of IBE system, which demonstrated secure encryption as indicated by indistinguishable definition. Another schemes were demonstrated secure under the standard oracle model but under selective security and adaptively secure.

2.) Attribute based encryption: Sahai and Waters defined ABE where complex access policies are expressed. Then Goyal, Pandey, Sahai[41]and Waters formulate two different ABE schemes-i.e. KEY policy ABE, Ciphertext policy ABE. In KP abe, attributes are attached with key which are distributed for decryption of cipher text for allowing decryption to only those data owned by particular party. In cipher text policy based abe, policy for allowing decryption or learning functions are attached to cipher text. In both the cases, if the policy is satisfied, then only party is allowed to decrypt ciphertext.

6.3 Predicate Encryption without Index

Predicate without public index is useful when faster results are needed without publicizing index associated with data. While above scheme take into account expressive types of accessing mechanism, they are restricted in following points. Firstly, the index associated is a part of the empty function-ality which is given clearly, this itself is private information. Secondly, computation on data which is encrypted is not allowed. Following is predicate encryption system that does not leak index.

1.)Inner Product Based Encryption: Katz, Sahai and Waters[41] proposed framework for testing if dot operation over the ring Zn is equivalent to 0, where N is a result of three arbitrary prime picked by the setup PPT algorithm. Inner product operation has proven to be faster approach in certain applications of functional encryption. In this approach, vectors of key, vector of random value and vector of data have to be at 90 degree to each other for dot product to be 0. After that, Okamotoand and Takshima gave development over the field Fp.

7. EVALUATION OF PRIVACY PRESERVI-NG ALGORITHM

Deciding evaluation criteria for assessing privacy preserving data mining is itself an issue of research. Identifying application of privacy and deciding evaluation criteria for algorithm is a critical matter. It is a frequent case that no security protecting calculation exists that beats all the others on all conceivable criteria. Maybe, a calculation may perform better than another on particular criteria, for example, execution and/or information utility. It is therefore essential to give clients an arrangement of measurements which will empower them to choose the most suitable security protecting procedure for the current information, concerning some particular parameters required for improving.

A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms, is given below [3]:

- The *performance* of the proposed calculations as far as time is concerned would rely on the time required by every calculation to conceal a predefined arrangement of touchy data and correspondence time between synergistic gatherings which ought to be considered.
- The data utility after the application of the privacy preserving technique, which is equivalent with the minimization of the information loss or else the loss in the functionality of the data; In case of non cryptography approaches information loss is higher compared to cryptography approach. Sanitized database and original database differs in variance result in clustering which can be the base of information loss.
- The *level of uncertainty* with which the sensitive information that have been hidden can still be predicted; The algorithm which exhibits maximum uncertainty can be preferred for privacy preserving data mining.
- The *resistance* accomplished by the privacy algorithms, to different data mining techniques. So as to accommodate a complete assessment of sanitization algorithms, we have to gauge its continuance against information mining procedures which are not the same as the procedure that a disinfection calculation has been created for. We call such parameter the transversal perseverance.

Table 6. Functional Encryption Schemes

Author	Year	Approach	Findings
Naveed, Shashank, Manoj[39]	2014	Controlled functional encryption, Inner product construction	Framework for data privacy based on cryptography assumptions
Shashank, Shweta[40]	2013	IND and SIM based FE definition, data privacy	Selectively secure FE against FE
Allison, Tatsuaki[46]	2011	KP & CP based ABE, bi - linear map, Inner Product based	Fully secure in dual encryption method
Barbosha, Farshim[42]	2012	Verifiable computing, strong secure DHE	Works in non-adaptive, bounded environment, it can't provide function privacy
Sergey, Vinod, Hoeteck[43]	2012	FE with Bounded collusion via Multiparty Computation	Q bounded, non adaptive, selective secure

8. CONCLUSIONS

The paper includes each and every technique on privacy preserving data mining. Till now, researchers are working on standardize model for privacy preserving data mining. Big data has come into picture and privacy concerns have grown more for data to be secure for individual's privacy.

Privacy preservation can be applied to certain limit based on data mining algorithm. Privacy and accuracy is a pair of contradiction. One overrides the result of another. In distributed environment, we concern about finding balance between algorithm complexities, computational cost and security. Many data mining algorithms are proposed so far, but not a single algorithm is up to the mark. Researchers need to combine advantages of all the algorithm and make a general model for privacy preserving data mining. Many proposed architectures of ppdm exhibits higher amount of cost. Against that, they provide higher measure of privacy and security. They are secure against security attacks like CCA2, CPA. They also secure against privacy attacks like inconsistent shares with honest party, consistent shares with adversarial party. Current survey is based on many cryptography assumptions. In future, experiments consisting of more number of parties would be the focus, with large amount of data tested under real world scenarios.

9. REFERENCES

[1]. Ontario. Information and Privacy Commissioner, and Ann Cavoukian.Data mining: Staking a claim on your privacy. 1997.

[2]. Liu Yu, Dap eng L, et al, Survey of research on anonymilization technology in data publication, Computer Application, pp. 2361-2364, 2009.

[3]. Verykios, Vassilios S.,et al."State-of-the-art in privacy preserving data mining."ACM Sigmod Record 3.1(2004): 50-57.

[4]. R. Agrawal and S. Ramakrishanan, Privacy preserving data mining ACM sigmod record, 2004.

[5]. Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao-Kui. Privacy Preserva- tion in Database Applications: A Survey.C hinese journerl of computer,2009

[6]. Yan Zhao1 Ming Du2 Jiajin, Le1 Yongcheng Luo1, A Survey on Privacy Preserving Approaches in Data Publishing. First International Workshop on Database Technology and Applications, 2009

[7]. A. Shamir. How to share a secret. Communications of the ACM,22(11):612 613, November 1979.

[8]. A. C. Yao. Protocols for secure computations(extended abstract). In 23rd Annual Symposium on Foundations of Computer Science. IEEE, 1982.

[9]. J.Benaloh, Dense probabilistic encryption. Citeseer .ist . psu.edu /benaloh94dense.html,1994.

[10]. Oliveira, Stanley RM, and Osmar R. Zaiane. "Privacy preserving frequent itemset mining." Proceedings of the IEEE international con- ference on Privacy, security and data mining-Volume 14. Australian Computer Society, Inc., 2002.

[11]. P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In Advances in Cryptology EUROCRYPT'99, pages 223-238. Springer, 1999.

[12]. Xiaolin Z. and Hongjing B. Research on privacy preserving classifica- tion data mining based on random perturbation. National conference of Information. Vol 1. No 1.,2010.

[13]. Kamakhi P. and Vinnaiya babu. Preserving privacy and sharing the data using classification on perturbed data. IJSCE. Vol 2. No 3. 2010.

[14]. Jagannathan, Geetha, and Rebecca N. Wright. "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data." Pro- ceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

[15]. Kantarcioglu, Murat, and Chris Clifton. "Privacy preserving distributed mining of association rules on horizontally partitioned data." IEEE Transactions on Knowledge and Data Engineering 16.9(2004) : 1026-1037.

[16]. Patel, Sankita, Sweta Garasia, and Devesh Jinwala. ”An Efficient Approach for Privacy Preserving Distributed K-Means Clustering Based on Shamirs Secret Sharing Scheme.” Trust Management VI. Springer Berlin Heidelberg, 2012.

[17]. Kantarcoglu, Murat, Jaideep Vaidya, and C. Clifton. ”Privacy pre- serving naive bayes classifier for horizontally partitioned data.” IEEE ICDM workshop on privacy preserving data mining. 2003.

[18]. Vaidya, Jaideep, and Chris Clifton. ”Privacy-preserving k-means clus- tering over vertically partitioned data.” Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[19]. Samarati, Pierangela. ”Protecting respondents identities in microdata release.” Knowledge and Data Engineering, IEEE Transactions on 13.6 (2001): 1010-1027.

[20]. Duan, Yitao, and John F. Canny. ”Practical Private Computation and Zero-Knowledge Tools for Privacy-Preserving Distributed Data Mining.” SDM. 2008.

[21]. Friedman, Arik, Assaf Schuster, and Ran Wolff. ”k-Anonymous deci- sion tree induction.” Knowledge Discovery in Databases: PKDD 2006. Springer Berlin Heidelberg, 2006. 151-162.

[22]. Blanton, Marina. ”Achieving full security in privacy-preserving data mining.”Privacy, security, risk and trust (passat), 2011 ieee third international conference on social computing (socialcom). IEEE, 2011.

[23]. Yang, Bin, et al. ”Collusion-resistant privacy-preserving data min- ing.”Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[24]. Xu, Zhuojia, and Xun Yi. ”Classification of privacy-preserving distributed data mining protocols.” Digital Information Management (ICDIM), 2011 Sixth International Conference on. IEEE, 2011.

[25]. Fung, Benjamin CM, Ke Wang, and Philip S. Yu. ”Anonymizing classification data for privacy preservation. Knowledge and Data Engineering, IEEE Transactions on 19.5 (2007): 711-725.

[26]. Fang, Weiwei, and Bingru Yang. ”Privacy preserving decision tree learning over vertically partitioned data.” Computer Science and Software Engineering, 2008 International Conference on. Vol. 3. IEEE, 2008.

[27]. Dasseni, Elena, et al. ”Hiding association rules by using confidence and support.” Information Hiding. Springer Berlin Heidelberg, 2001.

[28]. Lin, Zhenmin, and Jerzy W. Jaromczyk. ”Privacy preserving two-party k-means clustering over vertically partitioned dataset.” Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on. IEEE, 2011.

[29]. Slavkovic, Aleksandra B., Yuval Nardi, and Matthew M. Tibbits. ”” Secure” Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases.” Data Mining Workshops, 2007. ICDM Work- shops 2007. Seventh IEEE International Conference on. IEEE, 2007.

[30]. Xiao, Ming-Jun, et al. ”Privacy preserving id3 algorithm over horizontally partitioned data.” Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on. IEEE, 2005.

[31]. Xiao, Ming-Jun, et al. ”Privacy preserving C4. 5 algorithm over horizontally partitioned data.” Grid and Cooperative Computing, 2006. GCC 2006. Fifth International Conference. IEEE, 2006.

[32]. Inan, Ali, et al. ”Privacy preserving clustering on horizontally partitioned data Data and Knowledge Engineering 63.3 (2007): 646-666.

[33]. Pang, Liaojun, et al. ”A verifiable (t, n) multiple secret sharing scheme and its analyses.” Electronic Commerce and Security, 2008 International Symposium on IEEE,2008.

[34]. Aggarwal, Charu C., and S. Yu Philip. ”A condensation approach to privacy preserving data mining.” Advances in Database Technology- EDBT 2004. Springer Berlin Heidelberg, 2004. 183-199.

[35]. Reza, M., and Somayyeh Seifi. ”Classification and Evaluation the PPDM Techniques by using a data Modification-based framework.” IJCSE, Vol3. No2 Feb (2011).

[36]. Pinkas, Benny. ”Cryptographic techniques for privacy-preserving data mining ACM SIGKDD Explorations Newsletter 4.2 (2002).

[37]. Pedersen, Thomas Brochmann, Ycel Saygn, and Erkay Sava. ”Secret charing vs. encryption-based techniques for privacy preserving data mining.” (2007)

[38]. Taylor, Ronald C. ”An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics.” BMC bioin- formatics 11.Suppl 12 (2010): S1.

[39]. Naveed, Muhammad, et al. ”Controlled Functional Encryption.” Pro- ceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, November-2014. Beimel, Amos, et al. ”Non-Interactive Secure Multiparty Computation.”Advances in CryptologyCRYPTO 2014. Springer Berlin Heidelberg, 2014. 387- 404.

[40]. Agrawal, Shashank, et al. ”Function Private Functional Encryption and Property Preserving Encryption: New Definitions and Positive Results.” IACR Cryptology ePrint Archive 2013 (2013): 744

[41]. Attrapadung, Nuttapong, and Benot Libert. ”Functional encryption for public-attribute inner products: Achieving constant-size ciphertexts with adaptive security or support for negation.” J. Mathematical Cryptology 5.2 (2012): 115-158.

[42]. Barbosa, Manuel, and Pooya Farshim. ”Delegatable homomorphic encryption with applications to secure outsourcing of computation.” Topics in CryptologyCT-RSA 2012. Springer Berlin Heidelberg, 2012. 296-312.

[43]. Gorbunov, Sergey, Vinod Vaikuntanathan, and Hoeteck Wee. ”Func- tional encryption with bounded collusions via multi-party computation." Advances in Cryptology CRYPTO 2012. Springer Berlin Heidelberg, 2012.162-179.

[44]. Boneh, Dan, Amit Sahai, and Brent Waters. ”Functional encryption: Definitions and challenges.” Theory of Cryptography. Springer Berlin Heidelberg, 2011. 253-273.

[45]. Yang, Xiaoyuan, Weiyi Cai, and Ping Wei. ”Multiple-authority-keys CP-ABE.”Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on. IEEE, 2011.

[46]. Lewko, Allison, et al. ”Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption.” Advances in Cryptology EUROCRYPT 2010. Springer Berlin Heidelberg,2010.62-91.