ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

Importance of Privacy Preservation in Very Large Database-based Datamining: Review

[I]Lekshmy P L, [II]Dr.Abdul Rahiman M
[I]Assistant Professor, Dept. of CSE, LBS Institute of Technology for Women, Kerala, India
IIPro Vice Chancellor, Dr. A P J Abdul Kalam Technological University, Kerala, India

Abstract

Privacy preservation is an important issue during data mining process. Privacy preservation plays an important role in many application domains. The current practices of data storage and data extraction are insufficient to avoid the revelation of data owner through basic search. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual secrecy. This may lead to excessive data distortion or insufficient protection. In this paper, the importance and different mechanism to achieve privacy preservation is studied. Privacy preservation has received considerable attention in research communities, and many approaches have been proposed for different level of security measures . In this survey, we systematically describe and evaluate different approaches to privacy preservation in data mining (PPDM), study the challenges in practical data extraction, clarify the differences and requirements that distinguish PPDM from other related problems, and propose future research directions.

Keywords

Privacy Preservation, Research, Data Owner, Data Mining, Requirements, Security

I. Introduction

Data mining and also knowledge discovery in databases are the two novel research fields which investigate the automatic extraction of previously unidentified patterns from huge amounts of data. New improvements in data collection, data dissemination and corresponding technologies have introduced a fresh generation of research where present data mining algorithms ought to be reviewed from a different viewpoint, this of privacy preservation. It's better documented that this new with no restricts explosion of new information over the Internet and other media, has hit to a level where threats against the privacy are really general on a periodical basis and also they get essential thinking.

Privacy preserving data mining [4], is a new research focus in data mining and statistical databases [13], where data mining algorithms are examined for the side-effects they obtain in data privacy. The important thought in privacy preserving data mining is two-fold. First of all, sensitive raw data such as identifiers, names, addresses and the corresponding, should be modified or reduced from the original database, consecutively for the recipient of the data not to be effective to compromise other person's privacy. Detailed person-specific data in its original form often contains sensitive information about individuals, and publishing such data immediately violates individual privacy. The current practice primarily relies on policies and guidelines to restrict the types of publishable data and on agreements on the use and storage of sensitive data.

Secondly, sensitive knowledge that can be mined from a database by utilizing data mining algorithms must too be omitted, since such knowledge can evenly better compromise data privacy, as it'll be pointed. The important aim in privacy preserving data mining is to produce algorithms for changing the original data in any way, thus the private data and private knowledge stay private still later the mining function. The trouble that rises when confidential information can be obtained from published data by unauthorized users is in addition normally addressed as the "database inference" trouble. This paper offers a classification and a wide description of the different techniques and methodologies that have been built in the field of privacy preserving data mining.

The paper follows as in section 2, a basic overview on privacy preservation technique is described, addressing the general aspects and their relationships during data mining process. Section 3 demonstrates the several related works on privacy preservation necessities that have to be attained in data mining. Section 4 examines the critical issues in privacy to attain the securities necessities earlier presented whereas section 5 defines the most important issues of privacy preservation. At last, section 6 summarizes the major conclusions and lessons acquired from this study and shows future investigate directions on securities.

II. Classification of Privacy Preserving Techniques

There are numerous approaches that have been taken for privacy preserving data mining. They are categorized based on the following properties:

Data Distribution

The first property denotes the distribution of data. Some of the approaches are built for centralized data, whereas others concern to a distributed data situation. Distributed data situations could also be categorized into horizontal data distribution and also vertical data distribution. Horizontal distribution mentions these cases where several database records exist in various places, whereas vertical data distribution denotes to the cases where all the values for several attributes exist in various positions.

Data Modification

The second property relates to the data modification strategy. Generally, data modification is employed to change the original values of a database that requires to be released to the public and in this manner to guarantee high privacy protection. It's significant that a data modification method must be together with the privacy policy taken by an organization. Methods of modification contain:

- perturbation, that is established by the modification of an attribute value by a fresh value (i.e., altering a 1-value to a 0-value or adding noise)
- blocking, is the substitution of a present attribute value by a "?"

- aggregation or merging is the mixture of different values into a lower class
- swapping that denotes substituting values of individual records
- sampling mentions publishing data for solely a sample of a population.

Data Mining Algorithm

The third property denotes to the data mining algorithm, for which the data modification is happening. This is really something that's not identified earlier, however it helps the examination and invention of the data hiding algorithm. The difficulty of hiding data for a mixture of data mining algorithms is taken for future research agenda. For the present, several data mining algorithms are taken in separation of each other. Together with them, the foremost essential thoughts are produced for classification data mining algorithms, such as decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

Data or Rule Hiding

The fourth property mentions whether raw data or aggregated data must be concealed. The complexness for concealing aggregated data in the kind of regulations is certainly high and for this cause, generally heuristics are produced. The decreasing of the amount of public information makes the data miner to get weaker inference regulations that won't permit the inference of secret values. This procedure is too called as "rule confusion".

Privacy Preservation

The last property that is the foremost essential that denotes the privacy preservation method utilized for the selective alteration of the data. Selective alteration is needed to attain high usefulness for the altered data given that the privacy isn't threatened. The methods that are used for this cause are:

- heuristic-based methods such as adaptive modification which alters solely chosen values that reduce the usefulness loss instead of all obtainable values
- cryptography-based methods such as secure multiparty calculation where a calculation is protected if at the terminal of the calculation, no party knows anything apart from its own input and the results, and
- reconstruction-based methods where the original distribution of the data is rebuilt from the randomized data.

It is significant to recognize that data modification results in degradation of the database performance. To measure the degradation of the data, two metrics are chiefly utilized. The first one evaluates the confidential data protection, whereas the second calculates the loss of functionality.

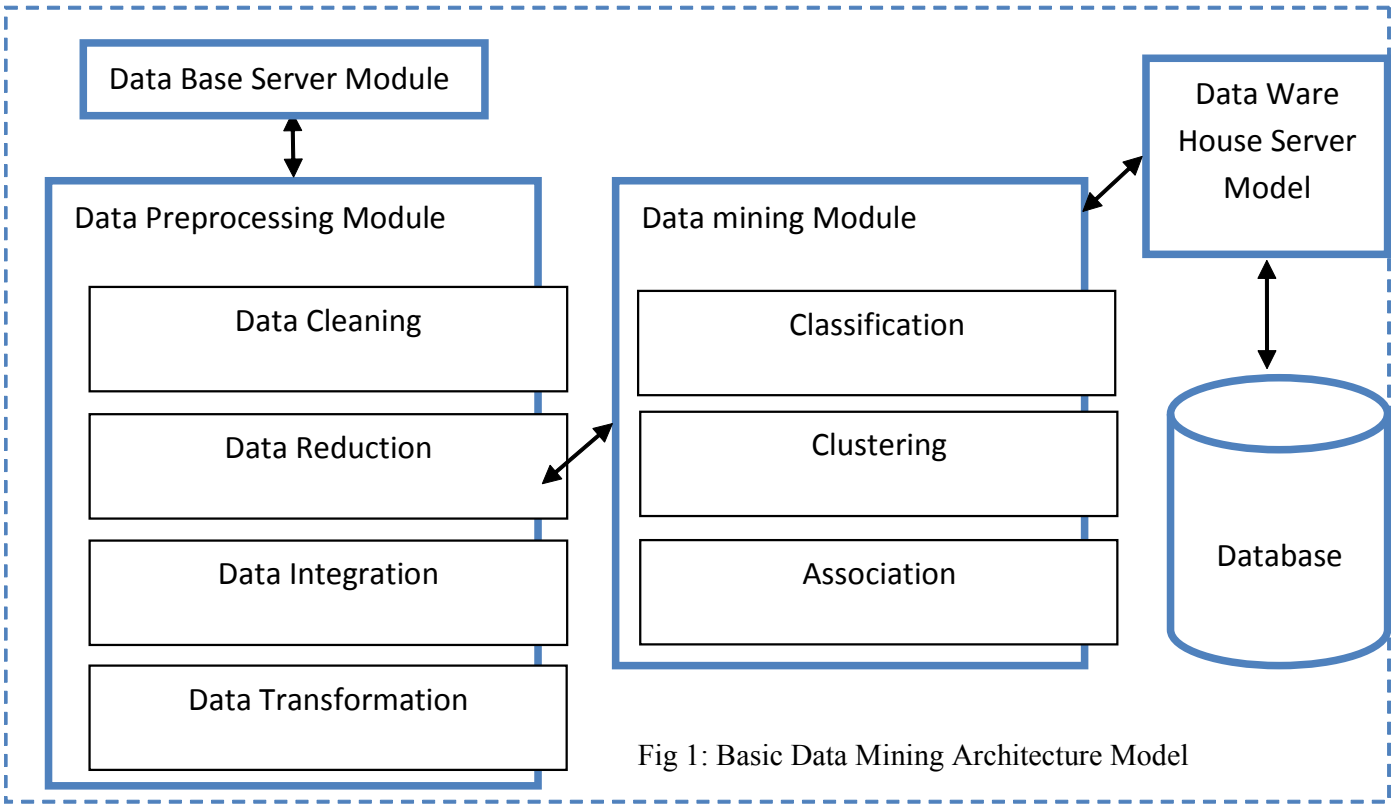

Fig 1: Basic Data Mining Architecture Model

III. Related Works

Barni et al. [1] propose a framework that to attain the twofold goal of allowing the processing of biomedical signals while ensuring the privacy of the signal's owner. The privacy-preserving classification of electrocardiogram (ECG) signals in the semi-honest model is considered. A protocol implementing the functionality is developed by resorting to generic secure two-party computation (STPC) techniques permitting two parties to compute the output of a public function on their relevant private inputs. In this framework, first, it presents two SSP (secure signal processing) protocols for the classification of ECG signals, built upon the classification algorithm. In particular, the SSP classifier relies on a subset of the features and by either a quadratic discriminant function (QDF) classifier, or by a neural network (NN). The former protocol used to better the overall efficiency of the protocol. The second protocol is totally new and relies

on the protocol for secure NN computation. Both protocols are designed by merging the advantages provided by homomorphic encryption (HE) and garbled circuits (GC). It also use the additively homomorphic cryptosystem of Paillier. GCs are an efficient method for SFE of Boolean circuits. The encryptions/garblings here can't be operated on directly, but need helper information which is generated and exchanged in the setup phase in form of a garbled table for each gate. The framework for modular design of efficient two-party secure function evaluation (SFE) protocols also included. This permits to modularly composing SFE protocols as a sequence of operations on encrypted data.

Camara et al. [3] offer a scheme to determine global informative gene subset to get the highest cancer classification accuracy, with limits on sharing of information. They introduce a novel RFE-SVM (recursive feature elimination- support vector machine) method. RFE-SVM algorithm is a wrapper feature selection method which generates the ranking of features utilizing backward feature elimination. It was originally planned to execute gene selection for cancer classification. Its basic idea is to reject redundant genes and yields better and more compact gene subsets. The features are eliminated corresponding to a criterion related to their support to the discrimination function and the SVM is re-trained at each step. RFE-SVM is weight-based method, at each step the coefficients of the weight vector of linear SVM are employed as the feature ranking criterion.

RFE-SVM has four steps. To train an SVM on the training set, order features utilizing the weights of the resulting classifiers, eliminates features with the smallest weight and repeat the process with the training set limited to the remaining features. The additive homomorphic encryption and decryption scheme are used to delineate RFE-SVM. This protects the privacy of patients while preserving researchers' ability to analyze globally specific genes. It increases the computational complexity and low performance.

Nix et al. [14] proposes a mechanism to encourage truthful data sharing for distributed data mining. Two methods are developed which does not need the ability to audit or determine the data, one for the non-cooperative case and one for the cooperative case. A Vickrey-Clarke-Groves (VCG) mechanism based on the accuracy of the result itself in order to encourage correct data reporting is introduced. Also another mechanism based on the Shapley value that encourages truthful sharing in the cooperative setting is offered. Classification tasks can be chosen for three reasons. First, it's widely accepted measure of utility, i.e., classification accuracy. Then used in association rules mining, recommender systems and countless other applications. At Last, the results generalize better to any task with a well-formed accuracy and utility metric.

Nix et al. describe a game scenario outlining the process for some data mining task. This model is a simple model in which a mediator performs the data mining computations. The mediator can be removed using the cryptographically secure techniques. Then the assured information sharing game in the context of a cooperative game is defined using a very simple method. The calculation of the Shapley value can be parallelized utilizing cloud architecture, as each model computation is independent of the others. Naïve Bayes classification algorithm can be implemented quite efficiently, without employing homomorphic encryption, with random data hiding techniques. This is incentive compatible for distributed data mining and increases the overall accuracy. However, computing cost is increased.

Daniel and Ashwin [6] urge a framework that preserves privacy of an individual and utility without making supposals about how the data are generated. No-free-lunch theorem that defines non-privacy as a game is used. To catch the notion of utility, it defines a lower bound on utility known as the discriminant. It measures whether there exist any query that can be answered reasonably accurately. It rejects the use of auxiliary side information, introduced a more natural notion of utility, and gave a shorter and transparent proof. In contrast to the forest fire and copying models, MVS model is really an ergodic Markov chain (whose state space comprises of graphs) and has a steady state distribution. Thus the number of edges doesn't have a tendency to increase over time as with the other two models. MVS is a continuous time process. This scheme protects against attackers. Still need improvement in privacy for correlated data.

Sramka et al. [16] introduce an approach for data mining utility that can be adjusted to capture the needs of data users. They propose a candidate sanitization method and define utility of a mining algorithm when applied to sanitized data in terms of the predictions that it can make about the original data. They model the utility of data users utilizing weights for records in the database to present the relative importance of the record and an appropriate error function to indicate the cost of making errors in predictions. This is called good utility. The utility function can also be used to express utility of an adversary by measuring correct predictions of sensitive attributes for individuals in the database. This is called bad utility or anti-utility.

A sanitization mechanism also developed which is a privacy-preserving algorithm that transforms and releases the original data that includes sensitive values and relations in a way that preserves and protects the privacy of these sensitive values and relations. The core component of the framework contains data mining algorithms. They are used for measuring and establishing both usefulness and privacy. It use the data mining utility to capture both the demands of mining users and intentions of adversaries.

Martin et al. [12] offers a scheme to measure the amount of disclosure of sensitive information in the worst case. The given approach begins with modeling the data publishing situation. Then two popular sanitization methods are defined. First, bucketization, that to partition the tuples into buckets and then to separate the sensitive attribute from the non-sensitive ones at random permuting the sensitive attribute values within each bucket. The sanitized data then comprises of the buckets with permuted sensitive values. Second is full-domain generalization, where it coarsens the non-sensitive attribute domains.

The sanitized data includes the coarsened table along with generalization used. To get identification information is to link quasi-identifying non-sensitive attributes released in the bucketization with publicly available data. In order to measure the power of knowledge, it employs the notion of a basic unit of knowledge and it offers a language which consists of finite conjunctions of such basic units. It provides a polynomial time algorithm for calculating the maximum disclosure even when the attacker has knowledge of dependencies. A polynomial-time dynamic programming algorithm for calculating the maximum disclosure is also introduced. This approach efficiently computes the maximum disclosure for any given bucketization. This method leads to maximum disclosure to sets of simple implications. Also suffers from an exponential blowup in the number of basic units required to express arbitrary DNF formulas. Still it has the complexity problem.

Roth and Roughgarden [15] present a scheme that answers arbitrary predicate queries that arrive online. They propose the

median mechanism. The median mechanism is a new method with a non-trivial utility guarantee and poly logarithmic privacy cost, even in the non-interactive setting. It designs an indicator vector method to privately release an indicator vector which distinguishes between hard and easy queries online. Easy queries are answered employing the corresponding median value; hard queries are answered as in the Laplace mechanism. It uses the random walk technology for convex bodies to perform efficient random sampling. This can be sampled utilizing the grid walk. Estimation error instead of brute force is easily controlled via the Cherno bound. Small uniformly sampled databases would also be good for the efficient version of this mechanism. DatabaseSample procedure generates databases with high probability, is close to uniform. It chooses a probability distribution uniformly at random from the positive quadrant. This scheme guarantees privacy. The basic implementation of the median mechanism is not efficient. This is not useful for all queries.

Bhaskar et al. [2] propose a framework that accurately discover and release the most significant patterns along with their frequencies in a data set comprising sensitive information. They present two efficient algorithms for discovering frequent itemsets in sensitive data sets, namely, Exponential mechanism based algorithm and Laplace mechanism based algorithm, both satisfy differential privacy. Differential privacy bounds the effect that any single record is on the distribution of the published information. In the exponential mechanism based algorithm, along with the notions of true frequency and noisy frequency, it has a notion of a truncated frequency. The analysis of privacy relies on bounding the sensitivity of the truncated frequency. This has the steps such as preprocessing, sampling and perturbation. In the perturbation step, it utilizes the Laplace noise addition technique independently on the true frequencies of the K itemsets selected in the sampling step. Composition lemma is applied on the sampling and the perturbation step together to ensure differential privacy. Laplace mechanism based algorithm use the idea to add independent Laplace noise to the frequencies of all itemsets and choose the K itemsets with the highest perturbed frequencies. This scheme guarantees the performance. But it has small additive error. It cannot provide the exact solution to the FIM problem.

Wong et al. [17] urge an approach to protect both identifications and relationships to sensitive information in data. They propose an (α , k)-anonymity model where α is a fraction and k is an integer. k-anonymity property is focused. The k-anonymity model assumes a quasi-identifier, which is a set of attributes that may function as an identifier in the data set. It extend Incognito, a global-recoding algorithm for the k-anonymity problem, to solve the problem for (α , k)-anonymity. Both parameters α and k are intuitive and usable in real-world applications.

Parameter α caps the confidence of implications from values in the quasi-identifier to the sensitive value while the parameter k specifies the minimum number of identical quasi-identifications. Incognito algorithm is an optimal algorithm for the k-anonymity problem. A local-recoding algorithm is also introduced to define a measurement for the amount of distortion generated by a recoding, which is called as the recoding cost. Then it offers a scalable local-recoding algorithm called top-down approach. The idea of the algorithm is to first generalize all tuples completely so that, all tuples are generalized into one equivalence class. Then, tuples are specialized in iterations. This approach is more scalable. It generates less distortion. In un-specialization step, the execution time is larger when α is smaller.

Fung et al. [7] offer a scheme that finds a k-anonymization, not necessarily optimal in the sense of minimizing data distortion that preserves the classification structure. They propose k-anonymization for classification. The data provider requires releasing a person-specific table for modeling classification of a specified class attribute in the table. Two types of information in the table are published. The first type is sensitive information and the second type is the quasi-identifier (QID). The quasi-identifier doesn't identify individuals but can be used to link to a person if the combination is unique. To construct a classifier, noises require to be generalized into patterns that are shared by more records in the same class. The data also comprises redundant structures. Then it suggests information metric to focus masking operations on the noises and redundant structures.

Lee and Mangasarian [9] urge a framework that generates a nonlinear kernel-based separating surface. They propose a Reduced Support Vector Machine (RSVM) that employs a randomly chosen subset of the data that is typically 10% or less of the original dataset to obtain a nonlinear separating surface. It utilizes a small random sample of given dataset as a representative sample with regard to the whole dataset in solving the optimization problem and in evaluating the nonlinear separating surface. It interpret this as a possible instance-based learning where the small sample is learning from the much larger training set by forming the appropriate rectangular kernel relationship between the original and reduced sets. SVM with a linear kernel is given by the quadratic program. If the classes are linearly inseparable, then two planes bound two classes with a soft margin decided by a nonnegative slack variable. This may be attributable to a decrease in data overfitting. The reduced dataset is all that is required in characterizing the final nonlinear separating surface. This is very important for massive datasets. RSVM is used in fraud detection. It does not affect test set correctness. But increases computational time.

Mangasarian et al. [11] introduce a scheme for a data matrix whose input feature columns are divided into groups belonging to different entities. They give a novel privacy-preserving support vector machine (PPSVM) classifier. PPSVM classifier for vertically partitioned data that is different from existing SVM classifiers and it relies on the two ideas. First, for a given data matrix, rather than using the usual kernel function it uses a reduced kernel. The second idea is that the columns of the random matrix are privately generated in blocks corresponding to the entities holding the blocks of input features. Every random column block of the matrix is established by solely one of the entities and that block of the random is known only to the entity that generated it and never made public. It makes utilize of the Hadamard separability of a nonlinear kernel which is satisfied by a Gaussian kernel. Then it introduces a linear and nonlinear privacy-preserving SVM classifier based on a privately generated and privately held random matrix by each entity. Each entity possesses a different set of input features used collectively to generate the SVM classifier. Employing a random kernel, a marked improvement of accuracy is acquired by the privacy-preserving SVM compared to classifiers generated by each entity utilizing its own data alone. This scheme does not reveal any of the privately-held data. Still has lower performance and exist computational complexity.

Huang et al. [8] urges a scheme for efficient, robust and automatic model selection for support vector machines (SVMs). They present a nested uniform design (UD) methodology. This method is applied to choose the candidate set of parameter groupings and carry out a *k*-fold cross-validation to measure the generalization

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

performance of each parameter combination. Contrary to conventional exhaustive grid search, this method can be addressed as a deterministic analog of random search. It first give a heuristic for setting up a two-dimensional search box in the parameter space that is capable to automatically scale the distance factor in the Gaussian kernel. Apart from the search scheme, it's always important to set up a proper search region. Once the search region is established, it applies the 2-stage UD methodology to choose the candidate set of parameter combinations and execute a *k*-fold cross-validation to measure the generalization performance of each parameter combination.

The 2-stage uniform design method first sets out a crude search for a highly likely candidate region of global optimum and then confines a finer second-stage search in that. For the upper and lower bound estimations based on a random subset it suggests the steps like, first, randomly sample a small subset from the entire data set; next compute the upper and lower bounds utilizing this random subset and finally adjust the bounds by a multiplicative factor. It delineates this search strategy based on nested UDs to decrease the number of trials in parameter combinations. It is always significant to set up a proper search region. It offers a heuristic for determining the search range which is able to automatically scale the distance factor in the Gaussian kernel. It also used a stratification scheme in dividing the entire data set to preserve the similarity pattern between training and test data sets. It reduces number of parameter trial and also provides the flexibility. Computing time is higher for some regions. One should be cautious for using high dimensional UDs because scattering points becomes sparse.

Liu et al. [10] introduce an approach that provides an effective and efficient balance between data utilities and privacy protection beyond its fast run time. They present a class of novel privacy-preserving data distortion methods in the collaborative analysis situations based on wavelet transformation also offer a new privacy breach algorithm in the collaborative analysis that can threaten the data privacy, yet with the distorted data values, in the single basis wavelet transformation case. Wavelet transformation is widely used in signal processing and noise suppression. A multi-basis wavelet data distortion strategy for improved privacy preserving is also presented. Then it employ discrete wavelet transformation (DWT) to simultaneously decompose the original data into approximation coefficients and detail coefficients and then suppress the high frequency detail coefficients to achieve data distortion. Such wavelet decomposition process is applied recursively with high and low passing filters on the approximation coefficients and then down-sampled. Inverse discrete wavelet transformation (IDWT) on approximation coefficients and suppressed detail coefficients to transform back the dataset to keep the same dimension as the original dataset is then introduced. The advantage is it can keep a better privacy level by preventing the intruders from exploiting the data. But breach process will not be effective in the multi-basis wavelet distortion method. It is difficult to adjust individual threshold values to get the best analysis result.

IV. Critical Design Issues

The design difficulty that affects the overall performance of privacy preserving data mining based on the earlier work research contains:

Efficiency

The overall processing of PPDM is employed to recognize the effective of the scheme. The capableness of a privacy preserving algorithm is to run the approach with good performance utilizing all the resources involved by the algorithm. The effectiveness consists of the amount of data extraction from the specified size of database. This offers the overall performance of the scheme.

Scalability

The amount of data to be mining without exposing the secret information of data owner denotes as scalability of PPDM. With increase size of information PPDM should efficient to extract information that measures the capability direction of a PPDM algorithm for raising sizes. The low speedy is the fall in the effectiveness of a PPDM algorithm for maximizing data dimensions; the best is its scalability. Thus, it's essential to measure the scalability of a PPDM technique as an evenly significant demand of such a type of algorithms.

Data Quality

The essential characteristic of PPDM contain Data Quality that is named as the common measure evaluating the position of the individual items comprised within the database following the implementation of a privacy preserving technique. Almost of privacy preserving algorithms typically use perturbation techniques to sanitize data from secret information (both private data items and also the complex data correlations). The data quality is, therefore, a significant parameter to consider in the estimation of a PPDM technique. The application of a privacy preserving technique, taken both as the quality of data themselves and also the quality of the data mining results later the hiding strategy.

The data quality throughout the implementation of privacy preserving algorithm is fundamentally based on

- **Accuracy:** it evaluates the finalization of a sanitized value to the original value as a result.
- **Completeness:** it offers the degree of dropped data in the precompiled database.
- **Consistency:** it's connected to the internal condition of PPDM that build the relationships which should maintain among several areas of a data item or among data items in a database.

Hiding Failure

This feature normally denotes to the capacity of keeping fun successful result that mostly present because of non-efficient feature extracting privacy preserving algorithm. Almost of the built up privacy preserving algorithms are planned with the aim of getting zero hiding failure. Therefore, they conceal all the patterns taken secret. The majority of privacy preserving algorithm selects the amount of secret data which must be concealed so as to determine a balance between privacy and knowledge discovery.

Degree of Privacy

This contains the general level of privacy offered by the scheme provided by a privacy preserving technique that too calculates the degree of uncertainty. Based on the degree of sensitivity hidden information may be expected by the system easily. The degree of privacy might be presented as, if the perturbed value of an attribute may be calculated with a confidence, that belong to an interval [a, b], and then the privacy is calculated by (b-a) with confidence c. This information doesn't function better since it doesn't take the distribution of the original data along with the perturbed data.

V. Significant Research Problems

Privacy preserving data mining plays an important role in data retrieval process. PPDM is dependent on numerous aspects like clustering, classification and anonymity, amount of dataset, multidimensionality and delay of processing, cryptographically approaches and more which affects the overall performance of the system. From the study of previous available privacy preserving approaches, the most researchable area includes classification and anonymity of data during data extraction process.

A. Classification Approaches

The classification approach plays an important role during privacy preservation approach. Classification plays an important role during the storage process of the data in the system. In order to obtain sensitive information to be hidden, the database is essentially modified through the insertion of false information or through the blocking of data values. But, these approaches don't really offer improved security that might not guarantee the scalability of data sets. Hence to increase the degree of privacy new and efficient protocols have to be designed.

B. Privacy Preserving Approaches

Preservation of information related to the data owner is the prime motivation privacy preserving data mining process. In the study of privacy preservation approaches it was found that the available algorithms are not efficiently preserving the privacy at scalable system. The hidden information, however, can still be inferred even though with some uncertainty level. A sanitization algorithm then could be measured on the basis of the uncertainty that it establishes during the reconstruction of the hidden information. Hence more research is necessary to ensure the degree of privacy and ability of approaches to reduce hiding failure information.

VI. Conclusion

Privacy preserving data mining (PPDM) can be considered as promising field for secure and efficient data hiding approaches. It was identified that efficient classification procedure is necessary for hiding the data owner identity. Classification mechanism will be used to reduce the dataset to masked information which can be later used for the retrieval of information. The inability to generalize the results for classes of categories of data mining algorithms might be a tentative threat for disclosing information. As explained an efficient privacy preserving algorithm is highly recommended to increase the degree of security and hiding failures mechanisms. In spite of above discussed open research problems, we believe that PPDM will be one of the most promising technologies for next-generation data analytics.

References

[1] *Barni M., Failla P., Lazzeretti R., Sadeghi A.-R., and Schneider T., "Privacy-Preserving ECG Classification With Branching Programs and Neural Networks," Information Forensics and Security, IEEE Transactions, vol.6, no.2, June 2011, pp. 452-468.*

[2] *Bhaskar R., Laxman S., Smith A. et al., "Discovering frequent patterns in sensitive data," Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, 2010, pp. 503–10.*

[3] *Camara F. et al, "Privacy Preserving RFE-SVM for Distributed Gene Selection," IJCSI, vol. 9, issue 1, no. 2, 2012, pp. 154-159.*

[4] *Chris Clifton and Donald Marks, "Security and privacy implications of data mining," In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996, pp. 15–19.*

[5] *Daniel E. O'Leary, "Knowledge Discovery as a Threat to Database Security," In Proceedings of the 1st International Conference on Knowledge Discovery and Databases, 1991, pp. 107–516.*

[6] *Daniel Kifer and Ashwin Machanavajjha, "No Free Lunch in Data Privacy," SIGMOD'11, 2011, pp. 193-204.*

[7] *Fung B.C.M., Wang K., and Yu P.S., "Anonymizing classification data for privacy preservation," IEEE Trans Knowledge Data Eng., vol. 19, 2007, pp. 711–25.*

[8] *Huang C.-H., Lee Y.-J., Lin D.K.J., and Huang S.-Y., "Model selection for support vector machines via uniform design," In Machine Learning and Robust Data Mining of Computational Statistics and Data Analysis, Amsterdam, Elsevier PublishingCompany,2007. http://dmlab1.csie.ntust.edu.tw/downloads/papers/UD4SVM013006.pdf.*

[9] *Lee Y.-J., and Mangasarian O. L., "RSVM: Reduced support vector machines," In Proceedings First SIAM International Conference on Data Mining, Chicago, April 5-7, 2001. ftp://ftp.cs.wisc.edu/pub/dmi/tech- reports/0007.pdf.*

[10] *Liu L., Wang J., Lin Z., and Zhang J., "Wavelet-based data distortion for privacy-preserving collaborative analysis," Technical Report 482-07, Department of Computer Science, University of Kentucky, Lexington, KY 40506, 2007. http://www.cs.uky.edu/jzhang/pub/MINING/lianliu1.pdf.*

[11] *Mangasarian O. L., Wild E. W., and Fung G. M., "Privacy-preserving classification of vertically partitioned data via random kernels," Technical Report 07-02, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, Sep. 2007.*

[12] *Martin D.J., Kifer D., Machanavajjhala A., et al., "Worst-case background knowledge for privacy-preserving data publishing," IEEE 23rd International Conference on Data Engineering, 2007, pp. 126 –35.*

[13] *Nabil Adam and John C. Wortmann, "Security Control Methods for Statistical Databases: A Comparison Study," ACM Computing Surveys 21, no. 4, 1989, pp. 515–556.*

[14] *Nix, Robert, Kantarciouglu, and Murat, "Incentive Compatible Privacy-Preserving Distributed Classification," Dependable and Secure Computing, IEEE Transactions, vol. 9, no. 4, 2012, pp. 451-462.*

[15] *Roth A., and Roughgarden T., "Interactive privacy via the median mechanism," Proceedings of the 42nd ACM symposium on Theory of computing. New York, NY, 2010, pp. 765–74.*

[16] *Sramka et al., "A Practice-oriented Framework for Measuring Privacy and Utility in Data Sanitization Systems," ACM, EDBT, Article No. 27, 2010.*

[17] *Wong R.C.W., Li J., Fu A.W.C. et al., "(a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia: ACM, 2006, pp. 754–9.*