International Journal of Computer Applications Technology and Research
Volume 6–Issue 2, 79-86, 2017, ISSN:-2319–8656

A Comprehensive Survey on Privacy Preserving Big Data Mining

S.Srijayanthi
Department of Computer Science and Engineering
R.M.K Engineering College, India

R.Sethukkarasi
Department of Computer Science and Engineering
R.M.K Engineering College, India

Abstract: In recent years, privacy preservation of large scale datasets in big data applications such as physical, biological and biomedical sciences is becoming one of the major concerned issues for mining useful information from sensitive data. Preservation of privacy in data mining has ascended as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation, and publishing. Privacy-Preserving Data Mining (PPDM) aids to mine information and reveals patterns from large dataset protecting private and sensitive data from being exposed. With the advent of varied technologies in data collection, storage and processing, numerous privacy preservation techniques have been developed. In this paper, we provide a review of the state-of-the-art methods for privacy preservation

Keywords: big data; confidential; data mining; privacy preservation; sensitive

1. INTRODUCTION

Privacy preservation in data mining has emerged as an unconditional prerequisite for exchanging private information in data analytics as internet phishing posed an intense menace on propagation of sensitive information over the web. Despite thriving of Big data provides potential values in healthcare, business analytics, government surveillance, and so on, substantial caution is essential in balancing the data utility and privacy preservation in the big data collection, storage and processing. Failing to protect privacy is immoral as it causes monetary loss and stern reputation impairment. Most methods for privacy computations use some form of data transformation to provide privacy preservation which reduce the granularity of representation resulting in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy.

A data set is viewed as a file with n records, where each record contains m attributes. The attributes can be classified as [16] identifiers, quasi-identifiers, confidential outcome attributes and non-confidential outcome attributes. There are several approaches implemented for privacy preserving data mining. They are classified based on the following dimensions [44]:

(i) Data modification
(ii) Data or rule hiding
(iii) Privacy preservation
(iv) Data mining algorithm
(v) Data distribution
(vi)

2. LITERATURE SURVEY

2.1 Data Modification

Data modification techniques modify the original values of a database and the transformed database is made available for mining.

The basic idea of value-based perturbation approach is to add random noise to the data values. The technique [3] proposed is based on random noise addition and is as follows: Consider n original data $A_1, A_{2,} \dots A_n$ of one-dimensional distribution following the same independent and identical distribution (i.i.d). The n random variables $B_1, B_2 \dots B_n$ are generated to hide Xi data values. Distributed data is generated as $W_1, W_2 \dots W_n$ where $W_i = A_i + B_i$. According to the perturbed dataset $W_1, W_2 \dots W_n$ and a reconstruction procedure based on Bayes rule, the density function will be estimated by Eq. (1)

$$f'_{X(a)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_{-\infty}^{a} f_Y(w_i - a)\, f_X(a)}{\int_{-\infty}^{+\infty} f_Y(w_i - z)\, f_X(z)\, dz} \qquad (1)$$

The reconstruction procedure is improved by Expectation Maximization (EM) algorithm. This method is able to retain privacy while accessing the information implicit in the original attributes. It is more effective in terms of information loss. The authors proved that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data.

A randomized Response (RR) technique was developed in the statistics community for the purpose of protecting surveyee's privacy and was first introduced by Warner [53]. Two models were proposed to solve the survey problem: Related-Question Model in which each respondent is asked two related questions, the answers to which are opposite to each other and Unrelated-Question Model in which two unrelated questions are asked with one probability for one of

the questions is known. The Multivariate Randomized Response (MRR) technique [18] was proposed for multiple-attribute data set. The method consists of two parts: the first part is the multivariate data disguising technique used for data collection; the second part is the modified ID3 decision tree building algorithm used for building a classifier from the disguised data. The framework [9] conducts a multivariate regression analysis to generate predicted probabilities for the sensitive item. They showed to use the sensitive attitude inferred from the multivariate regression analysis as a predictor for an outcome regression model.

The condensation based technique [4] was proposed to generate pseudo-data from clustered groups of k-records. Principal component analysis of the behaviour of the records within a group is used in the generation of pseudo-data. The use of pseudo-data provides an additional layer of protection. Also, the aggregate behaviour of the data is preserved thus useful for a variety of data mining problems. The technique [5] constructed groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. Then, pseudo-data were generated from each group to create a synthetic data set with the same aggregate distribution as the original data. Aggarwal [1] has proposed a method for anonymization of string data that creates clusters from the different strings, and then generates synthetic data which has the same aggregate properties as the individual clusters. Since each cluster contains at least k-records, the anonymized data is guaranteed to at least satisfy the definitions of k-anonymity.

The main idea of random rotation perturbation technique is that the original dataset with d columns and N records represented as X_{dXN}. The rotation perturbation of the dataset X will be defined as g(X) = RX . Where R_{dXd} is a random rotation orthonormal matrix. A key feature of rotation transformation is preserving the Euclidean distance, inner product in a multi-dimensional space. The optimal algorithm [11] perturbs all columns together. The authors defined an efficient multi-column privacy measure for evaluating the privacy quality of any rotation perturbation. The level of difficulty for the estimation of the original data is by variance of the difference. Let rij represent the $element(i, j)$ in the matrix R, and cij be the $element(i, j)$ in the covariance matrix of X. The Variance of Difference ith column is computed by Eq. (2)

$$Cov\left(X^{'} - X \right)_{(i,i)} = \sum_{d}\sum_{d} r_{ij} r_{ik} c_{kj} - 2\sum_{d} r_{ij} c_{ij} + c_{ii} \quad (2)$$

Geometric data perturbation consists of a sequence of random geometric transformations, including multiplicative transformation (R), translation transformation (Ψ), and distance perturbation (Δ). Authors Chen and Liu [12] have developed three protocols: (i) simple protocol to transmit encrypted perturbed data to the service provider. (ii) negotiation protocol enables multi-round voting to reach an agreed perturbation. (iii) The space adaptation protocol provides a better balance between scalability, flexibility of data distribution, and the overall satisfaction level of privacy guarantee. The authors Chen and Liu [13] proposed a multi-column privacy evaluation model and designed a unified privacy metric to address the problems. The authors analysed the resilience of the rotation perturbation approach against three types of inference attacks: naive-inference attacks, ICA-based attacks, and distance-inference attacks. The authors constructed a randomized optimization algorithm to efficiently find a good geometric perturbation that is resilient to the attacks.

Random projection projects a set of data points from a high dimensional space to a randomly chosen lower dimensional subspace. The basic idea of random projection arises from the Johnson-Lindenstrauss Lemma. The authors Kargupta et al. [23] proposed spectral filtering technique that can estimate values of individual data-points from the perturbed dataset and thus can be used to reconstruct the distribution of actual data as well. Signal-to-Noise Ratio (SNR) quantifies the relative amount of noise added to actual data to perturb it and is given by Eq. (3)

$$SNR = \frac{(Variance\ of\ \ Actual\ \ Data)}{(Noise\ Variance)} \quad (3)$$

The authors Liu et al. [32] showed that the projection can preserve the inner product, which is directly related to several distance-related metrics, by conducting row wise and column-wise projection of the sample data. The authors Li et al. [29] expanded scope of additive perturbation based PPDM to multi-level trust (MLT). The method allows data owners to generate differently perturbed copies of its data for different trust levels on demand, offering maximum flexibility to data owners. The key challenge lies in circumventing from combining copies at different trust levels to jointly reconstruct the original data. This is addressed by properly correlating perturbation across copies at different trust levels.

The Singular value decomposition SVD technique is used to distort portions of the datasets. The SVD of the data matrix A is given by the Eq. (4)

$$A = U \sum V^{T} \quad (4)$$

where A be a sparse matrix of dimension $n \times m$ representing the original dataset. U is $n \times n$ orthogonal matrix and V^{T} is $m \times m$ orthogonal matrix. A transformed matrix with a much lower dimension is defined by Eq. (5)

$$A_{k} = \bigcup_{k} \sum_{k} V_{k} T \quad (5)$$

The proposed data distortion method [48], sparsified SVD, is better than SVD. Entries smaller than a certain threshold in are set to zero after reducing the rank of the SVD matrices. This operation is called as dropping operation. The distorted data matrix $\overline{A}_k$ is written as Eq. (6)

$$\overline{A}_k = \overline{U_k} \sum_k \overline{V_k}^T \quad (6)$$

$\overline{A}_k$ is twice distorted in the sparsified SVD method and thus it is harder to reconstruct the entries in A. The computation of SVD for large scale dataset matrices is expensive which can be substantially reduced by employing clustered SVD strategies.

In the proposed algorithm [21], attributes are grouped according to their distance difference similarity by clustering the data set using decision tree classification. The algorithm packetizes the attributes in each group and for each group it creates an equivalence class following the unique attribute-distinct diversity anonymization model. The weights given to attributes improve clustering and give the ability to control the generalization's depth.

In Non-negative matrix factorization (NNMF) technique is a vector space method to obtain a representation of data using non-negative constraints. Considering $n \times m$ nonnegative matrix dataset A with $A_{ij} \geq 0$ and a pre-specified positive integer $k \leq \min\{n, m\}$, nonnegative matrix factorization finds two non-negative matrixes $W \in R^{n \times k}$ with $W_{ij} \geq 0$ and $H \in R^{k \times m}$ with $H_{ij} \geq 0$, such that $A \approx WH$ and the objective function given by Eq. (7) is minimized

$$f(W, \mathrm{H}) = \frac{1}{2} \|A - WH\|_F^2 \quad (7)$$

where $\|A - WH\|_F$ the Frobenius norm. Matrices W and H have desirable properties in data mining applications. The work in [51] contributed least-square compression form of original datasets and iterative methods to solve the least-square optimization problem.

Each release of the data must be such that every combination of values of quasi-identifiers can be indistinguishably matched to at least k respondents. In k-anonymity techniques [39], the granularity of representation of the pseudo-identifiers is minimized by generalization and suppression. k-Optimize algorithm [8] assumes ordering among the quasi-identifiers. The Incogito method [25] has been proposed for computing k-minimal generalization with the use of bottom-up aggregation along domain generalization hierarchies. [19] starts with a general solution, and then specializes some attributes of the current solution so as to increase the information, but reducing the anonymity. The reduction in anonymity is always controlled, so that k-anonymity is never violated.

A hybrid approach [49] proposed combined Top Down Specialization (TDS) and BUG (Bottom Up Generalization) together for efficient sub-tree anonymization over big data. The approach automatically determined which component to be used to conduct the anonymization when a data set was given, by comparing the user specified k-anonymity parameter with a threshold derived from the dataset. Both components TDS and BUG are developed based on Map Reduce (MR) to gain high scalability by exploiting powerful computation capability of cloud.

In TDS [50] scalable approach proposed, a data set is anonymized by performing specialization operations. In the first phase, data sets are partitioned and anonymized in parallel, producing intermediate results. The intermediate results are merged and further anonymized to produce k-anonymous data sets in the second phase. The goodness of a candidate specialization is measured by a search metric, Information Gain per Privacy Loss (IGPL).

BUG approach of anonymization is an iterative process starting from the lowest anonymization level. The goodness of a candidate generalization is measured by a search metric, Information Loss per Privacy Gain (ILPG).

ILPG of generalization is given by the Eq. (8)

$$ILPG(gen) = IL(gen) / (PG(gen) + 1) \quad (8)$$

Information loss is given by Eq. (9)

$$IL(gen) = \sum_{c \in Child(q)} \left(\left| \frac{R_C}{Rq} \right| I(R_c) - I(Rq) \right) \quad (9)$$

The privacy gain is given by Eq. (10)

$$PG(gen) = A_{p(gen)} - A_{c(gen)} \quad (10)$$

The extended k-Anonymity [47], (α, k)-Anonymity, combines two principles: (i) each equivalence class must have size at least k (ii) at most α percent of its tuples can have the same sensitive value. The authors presented an optimal global-recoding (α, k)-anonymization algorithm and a scalable local-recoding technique that shows less data distortion.

The k-Anonymity technique is vulnerable to many kinds of attacks if the background knowledge is known. Such kinds of attacks include are homogeneity attack and background knowledge attack. The l-diversity technique proposed not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes. Therefore, the l-diversity model [24] for privacy is defined as a group of indistinguishable tuples are l-diverse

if they contain at least l "well-represented" values for the sensitive attributes. Liu and Wang [31] proposed an extension of l-diversity using full-subtree generalization and suppression techniques. It is stated that the confidence of the adversary in inferring a target's sensitive information is bounded by the percentage $conf\left(S_i \mid QI_j\right)$ of the records that contain the same value Si in the equivalence class j. Authors limit this bound by guaranteeing $conf\left(S_i \mid QI_j\right) \leq \theta_i$, where parameter θi is a given privacy threshold in the interval [0, 1]. A dynamically created structure, Cut enumeration tree, enumerates all possible generalizations of QI attributes according to the generalization level and information loss of each candidate solution.

To prevent skewness attack, the authors [26] proposed a privacy model, called t-closeness, which states that an equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. It severely affects the data utility as it needs the distribution of sensitive values to be the same in all equivalence classes. The authors used Earth Mover Distance (EMD) to measure the closeness between two sensitive values which does not prevent attribute linkage on numerical sensitive attributes. Table 1. shows the comparison of data perturbation techniques.

Table 1. Comparison of perturbation techniques

Criteria	**perturbation**		
	Value Based	**Data Mining Task**	**Dimensional Reduction**
Privacy Loss	Average	Low	Very Low
Information Loss	Low	Very Low	Very Low
Modify Mining Algorithms	Yes	No	No
Data Dimension	Single	Multi	Multi

2.2 Association Rule Hiding

Association rule hiding algorithms can be divided into three classes namely: (i) Heuristic (ii) Border-based approaches (iii) Exact approaches

The heuristic approaches sanitize a set of transactions from the database to hide the sensitive knowledge. It is efficient and scalable. In several circumstances they suffer from undesirable side-effects that lead them to suboptimal solutions. The authors Atallah et al. [7] proposed a greedy iterative search algorithm to hide sensitive association rules through the reduction in the support of their generating itemsets. The limitation is the loss of support for a large itemset, as long as it remains frequent in the sanitized outcome. Verykios et al. [45] proposed two heuristic algorithms. The first algorithm hides the item having the maximum support from the minimum length transaction. The second algorithm sorts the generating itemsets with respect to their size and support. The algorithm removes the items from the corresponding transactions in a round-robin fashion, until the support of the sensitive itemset drops below the minimum support threshold. Amiri [6] proposed three effective, multiple rule hiding heuristics approach: (i) Aggregate approach (ii) Disaggregate approach (iii) Hybrid approach.

DSRRC (Decrease Support of Right hand side item of Rule Clusters) algorithm [34] clusters the sensitive rules based on certain criteria in order to hide as many as possible rules at one time. One shortcoming of this algorithm is that it cannot hide association rules with multiple items in antecedent and consequent. The authors Domadiya and Rao [15] introduced a heuristic based algorithm called Modified Decrease Support of RHS item of Rule Clusters (MDSRRC) to secure the delicate association rules using multiple items in consequent (RHS) and antecedent (LHS). This algorithm successfully addressed the drawbacks of rule hiding DSRRC algorithm.

Saygin et al [40] proposed the usage of unknowns instead of altering 1's to 0's and vice versa to hide association rules. The two schemes [46] proposed include unknowns and aimed at the hiding of predictive association rules. The algorithms proposed require a reduced number of database scans and exhibit an efficient pruning strategy. The first scheme decreases the confidence of a rule by increasing the support of the itemset in its LHS. The second approach reduces the confidence of the rule by decreasing the support of the itemset in its RHS.

The border based approach modifies the original borders in the lattice of the frequent and the infrequent patterns in the dataset. The sensitive knowledge is hidden by enforcing the revised borders in the sanitized database.

The Sun and Yu [41] proposed a scheme which first computes the positive and the negative borders in the lattice of all itemsets. A weight is assigned to each element of the expected positive border which is dynamically computed as a function of the current support. The algorithm deletes the candidate item that will have the minimal impact on the positive border. The authors Moustakides and Verykios [35] proposed an algorithm to remove all the sensitive itemsets belonging to the revised negative border. Among all minimum border itemsets, the one with the highest support is selected. This max-min itemset determines the item through which the hiding of the sensitive itemset will incur.

The exact approach considers the hiding process as a constraint satisfaction problem solved using integer or linear programming. Exact approaches are efficient than heuristic schemes, at a high computational cost. They formulate the sanitization process as a constraint satisfaction problem. The scheme [33] consists of an exact and a heuristic part in which the exact part formulates a Constraint Satisfaction Problem (CSP) with the objective of identifying the minimum number of transactions that need to be sanitized. An integer programming solver is then used to identify the best solution. An approach [20] uses the itemsets belonging in the revised positive and negative borders to identify the candidate itemsets for sanitization. It obtains efficient solution of the CSP, by using binary integer programming.

2.3 Privacy Preservation

It refers to the privacy preservation technique used for the selective modification of the data. The cryptographic methods tend to compute functions over inputs provided by multiple recipients without actually sharing the inputs with one another. The challenge is to conduct such a computation while preserving the privacy of the inputs. [14] presented four secure multiparty computation based methods that can support privacy preserving data mining. The methods described include, the secure sum, the secure set union, the secure size of set intersection, and the scalar product. The authors Kantarcioglu and Clifton [22] addressed the problem of secure mining of association rules over horizontally partitioned data based on the assumption that each party first encrypts its own itemsets using commutative encryption, then the ready encrypted itemsets of every other party. A secure comparison takes place between the final and initiating parties to determine if the final result is greater than the threshold plus the random value. Based on cryptographic techniques Chakravorty et al. [10], replaced the personal/quasi- identifiers of collected sensor data with hashed values before storing them into a de-identified storage. In the Dong and Chen [17] proposed an efficient secure dot product protocol based on the Goldwasser–Micali Encryption and Oblivious Bloom Intersection for privacy preserving association rule mining. The protocol is faster as it employs mostly cheap cryptographic operations, e.g. hashing and modular multiplication. The authors Wang et al. [52] proposed a privacy-preserving public auditing system for data storage security in cloud computing utilizing the homomorphic linear authenticator and random masking to guarantee that the third party would not learn any information stored on the cloud server during the auditing process.

The reconstruction based methods first randomize the original data to hide the sensitive data and then reconstruct the interesting patterns based on the statistical features without knowing true values. The work [3] addresses the problem of building a decision tree classifier from training data in which the values of individual records have been perturbed. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. The approach [2] is based on an Expectation Maximization (EM) algorithm for distribution reconstruction which converges to the maximum likelihood estimate of the original distribution on the perturbed data.

2.4 Data Mining Algorithm

To extract useful information from big data without breaching the privacy, privacy preserving data mining techniques have been developed to identify patterns and trends from data. These techniques can be broadly grouped into clustering, classification and association rule based techniques.

The authors Zhou et al. [54] proposed a parallel k-means clustering algorithm by using three functions of MapReduce. First, the Map tasks calculate the closest distance for data points from every initial centroid of clusters. Next, the combiner calculates a partial sum of values. The Reduce tasks compute the centroids by dividing the partial sum of samples in to the number of samples assigned to a similar cluster. The Mapper processes each data object and called several times which increases the problem in handling large data sets.

In Incremental k-means Algorithm (IKM) [36], the Mapper loads data segment, and executes the IKM on the loaded data segment. The Reducer receives the intermediate results and executes the IKM again to obtain the clustering results. This approach provides an approximate solution and does not provide exact clustering results. Li et al. [27] focused on concurrently running k-means processes based on MapReduce with multiple initial center groups. Its main objective is to avoid serial execution of k-means and more focus on initial centroids. In this approach, the hopeless k-means process attempts are abandoned, which speeds up the future iterations. However, because of using MapReduce, it still lacks the ability to cache data between iterations for improving performance.

Classification is a technique of identifying, to which predefined group a new input data belongs. Classification algorithm is designed to process data in two ways [42]. It either classifies the data by themselves or forward the input data to another classifier. It is computationally efficient particularly when handling large and complex data.

The algorithm [3] in which the original data are altered by adding random offsets was proposed. Bayesian formula is used to derive the density function of the original data. A random forest is built based on Mahout RF Partial implementation to classify imbalanced big data. The algorithm calculates the leaves weights for each tree. Then, the leaf weight is the accumulated weight divided by the

number of instances classified and then the algorithm combines the outputs from each mapper. For each instance in all classes, the accumulated weight is divided by the number of trees involved in the classification.

A global SVM classification model [43] was constructed based on gram matrix computation to securely compute the kernel matrix from the distributed data. The algorithm, [30] Privacy-Preserving SVM Classifier PPSVC approximates the decision function of the Gaussian kernel SVM classifier without compromising the sensitive attribute values possessed by support vectors. The PPSVC is robust against adversarial attacks and the accuracy is comparable to the original SVM classifier. Quantum based support vector machine [38] for big data classification minimizing the computational complexity and the required training data was proposed.

2.5 Distributed Privacy Preservation

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy.

The BOPPID (Boosting – based Privacy Preserving Integration of Distributed data) algorithm [28] in which each participant has different set of records with both common features and local unique attributes. AdaBoost algorithm was employed to build an ensemble classifier. By sharing the local models with each other, all the participators can build their individual integrated model without direct access to the datasets. To prevent "negative impact" during integration, the models from the other participators whose data distribution is very different from the data distribution of this participator are excluded. The proposed method overcomes the need of third-party and reduces the communication cost. An algorithm [37] was proposed for differentially private data release for vertically partitioned data. The two-party differentially private data release algorithm anonymized the raw data by sequence of specialization and added noise. The proposed distributed exponential mechanism takes candidate and score pairs as inputs. Candidates are selected based on their score functions. The score is determined using Max utility function given by Eq. 11

$$Max(D,v) = \sum_{c \in child(v))} \left(\max \left| D_c^{cls} \right| \right) \qquad (11)$$

3. CONCLUSION

The privacy preservation for data analysis is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. Each privacy preserving technique has its own importance. Data encryption and anonymization are widely adopted ways to combat privacy breach. However, encryption is not suitable for data that are processed and shared. Anonymizing big data and managing anonymized data sets are still challenges for traditional anonymization approaches. Privacy-preserving data mining is emerged for to two vital needs: data analysis in order to deliver better services and ensuring the privacy rights of the data owners. Substantial efforts have been accomplished to address these needs. In this paper, an overview of the recent approaches for privacy preservation was presented. The privacy guarantees, advantages and disadvantages and possible enhancement of each approach were stated.

4. REFERENCES

[1] C. Aggarwal, "On randomization, public information and the curse of dimensionality", In: IEEE 23rd International Conf. on Data Engineering, 2007.

[2] D. Agrawal, and C. Aggarwal. "On the design and quantification of privacy preserving data mining algorithms", In: Proc. of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2001.

[3] R. Agrawal, and R. Srikant, "Privacy-preserving data mining", ACM Sigmod Record. Vol. 29, No. 2, 2000.

[4] C. Aggarwal and S. Yu Philip. "A condensation approach to privacy preserving data mining", In: International Conf. on Extending Database Technology. Springer Berlin Heidelberg, 2004.

[5] C. Aggarwal and P. S. Yu, "On Variable Constraints in Privacy Preserving Data Mining", SDM. 2005.

[6] A. Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization", Decision Support Systems 43(1), pp.181-191, 2007.

[7] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios, "Disclosure limitation of sensitive rules" Knowledge and Data Engineering Exchange, 1999.(KDEX'99), Proc. 1999 Workshop on, pp. 45-52, IEEE, 1999.

[8] R.J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization" In: 21st International Conf. on Data Engineering (ICDE'05), pp.217-228, IEEE, 2005.

[9] G. Blair, K. Imai, and YY. Zhou. "Design and Analysis of the Randomized Response Technique", Journal of the American Statistical Association 110(511), pp. 1304-1319, 2015.

[10] A. Chakravorty, T. Wlodarczyk and C. Rong, "Privacy Preserving Data Analytics for Smart Homes" In Security and Privacy Workshops (SPW), pp. 23-27. IEEE, 2013.

[11]K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation", In: Fifth IEEE International Conf. on Data Mining (ICDM'05), pp. 4-pp. IEEE, 2005.

[12]K. Chen and L. Liu, "Privacy-preserving multiparty collaborative mining with geometric data perturbation", IEEE Transactions on Parallel and Distributed Systems 20(12), pp. 1764-1776, 2009.

[13]K. Chen, and L. Liu, "Geometric data perturbation for privacy preserving outsourced data mining", Knowledge and Information Systems 29(3), pp. 657-695, 2011.

[14]C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and MY. Zhu, "Tools for privacy preserving distributed data mining" , ACM Sigkdd Explorations Newsletter 4, no. 2, pp.28-34, 2002.

[15]N. H. Domadiya, U. P. Rao, "Hiding sensitive association rules to maintain privacy and data quality in database", In: Advance Computing Conf. (IACC), 2013 IEEE 3rd International, pp.1306-1310, 2013.

[16]J. Domingo-Ferrer, "A survey of inference control methods for privacy-preserving data mining", In: Privacy-preserving data mining, pp. 53-80. Springer US, 2008.

[17]C. Dong and L. Chen, "A fast secure dot product protocol with application to privacy preserving association rule mining", In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 606-617, Springer International Publishing, 2014.

[18]W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining", In: Proc. of the ninth ACM SIGKDD international conf. on Knowledge discovery and data mining, pp. 505-510, ACM, 2003.

[19]BCM. Fung, K. Wang, and PS. Yu. "Top-down specialization for information and privacy preservation, " In: 21st International Conf. on Data Engineering (ICDE'05), pp.205-216, IEEE, 2005.

[20]A. Gkoulalas-Divanis and V S. Verykios. "An integer programming approach for frequent itemset hiding", In: Proc. of the 15th ACM international conf. on Information and knowledge management, pp. 748-757. ACM, 2006.

[21]P. Jain, N. Pathak, P. Tapashetti, and A. S. Umesh. "Privacy preserving processing of data decision tree based on sample selection and Singular Value Decomposition", In: 9th International Conf. on Information Assurance and Security (IAS), 2013, pp. 91-95, IEEE, 2013.

[22]M. Kantarcioglu, and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data", IEEE transactions on knowledge and data engineering 16, no. 9, pp. 1026-1037, 2004.

[23]H. Kargupta, Hillol, S. Datta, Q. Wang, and K. Sivakumar. "On the privacy preserving properties of random data perturbation techniques." In: Third IEEE International Conference on Data Mining, 2003. ICDM 2003., pp. 99-106, IEEE, 2003.

[24]A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3, 2007.

[25]K. LeFevre, DJ. DeWitt, and R. Ramakrishnan. "Incognito: Efficient full-domain k-anonymity", In: Proc. of the 2005 ACM SIGMOD international conf. on Management of data, pp. 49-60, ACM, 2005.

[26]N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity", In: IEEE 23rd International Conf. on Data Engineering, pp. 106-115, IEEE, 2007.

[27] C. Li, Y. Zhang, M. Jiao, and Ge Yu. "Mux-Kmeans: multiplex kmeans for clustering large-scale data set", In: Proc. of the 5th ACM workshop on Scientific cloud computing, pp. 25-32, ACM, 2014.

[28] Y. Li, C.Bai, and CK.. Reddy. "A distributed ensemble approach for mining healthcare data under privacy constraints", Information sciences 330, pp. 245-259, 2016.

[29]Y. Li, M. Chen, Q. Li, and W. Zhang. "Enabling multilevel trust in privacy preserving data mining", IEEE Transactions on Knowledge and Data Engineering 24, no. 9, pp. 1598-1612, 2012

[30]KP. Lin, and MS. Chen. "On the design and analysis of the privacy-preserving SVM classifier", IEEE Transactions on Knowledge and Data Engineering 23, no. 11 pp. 1704-1717, 2011.

[31]Liu, Junqiang, and Ke Wang. "On optimal anonymization for l+-diversity", In: IEEE 26th International Conf. on Data Engineering (ICDE 2010), IEEE, 2010.

[32]K. Liu, H. Kargupta, and J. Ryan. "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining" IEEE Transactions on knowledge and Data Engineering 18(1), pp. 92-106, 2006.

[33]S. Menon, S. Sarkar, and S. Mukherjee. "Maximizing accuracy of shared databases when concealing sensitive patterns", Information Systems Research 16(3), pp. 256-270, 2005

[34]Modi, Chirag N., Udai Pratap Rao, and Dhiren R. Patel. "Maintaining privacy and data quality in privacy preserving association rule mining", In: International Conf. on Computing Communication and Networking Technologies (ICCCNT), 2010.

[35]GV. Moustakides and VS. Verykios, "A MaxMin approach for hiding frequent itemsets", Data & Knowledge Engineering 65(1), pp. 75-89, 2008.

[36]DT. Pham, SS. Dimov, and CD. Nguyen, "An incremental K-means algorithm", In: Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 218(7), pp. 783-795, 2004.

[37]N. Mohammed, D. Alhadidi, BCM Fung, and M. Debbabi, "Secure two-party differentially private data release for vertically partitioned data", IEEE Transactions on Dependable and Secure Computing 11, no. 1, pp. 59-71, 2014.

[38]P. Rebentrost, M. Mohseni, and S. Lloyd, "Quantum support vector machine for big feature and big data classification", arXiv preprint arXiv:1307.0471, 2013.

[39]P. Samarati, "Protecting respondents identities in microdata release", IEEE transactions on Knowledge and Data Engineering, 13(6), pp.1010-1027, 2001.

[40]Y. Sayginl, VS. Verykios, and C. Clifton. "Using unknowns to prevent discovery of association rules", Acm Sigmod Record 30, no. 4, pp. 45-54, 2001.

[41]X. Sun, and PS. Yu. "A border-based approach for hiding sensitive frequent itemsets", In: Fifth IEEE International Conf. on Data Mining (ICDM'05), pp. 8-pp. IEEE, 2005.

[42]C. Tekin and M. van der Schaar. "Distributed online big data classification using context information", In: 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2013, pp. 1435-1442, IEEE, 2013.

[43]J. Vaidya, H. Yu, and X. Jiang. "Privacy-preserving SVM classification", Knowledge and Information Systems 14, no.2, pp. 161-178, 2008.

[44]VS. Verykios, E. Bertino, IN. Fovino, LP. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", ACM Sigmod Record 33, no. 1, pp. 50-57, 2004.

[45]VS. Verykios, AK. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. "Association rule hiding", IEEE Transactions on knowledge and data engineering 16, no. 4, pp. 434-447, 2004.

[46]SL. Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules", In: IRI-2005 IEEE International Conf. on Information Reuse and Integration, Conference, 2005, pp. 223-228. IEEE, 2005.

[47]RCW. Wong, J. Li, AWC. Fu, and K. Wang. "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing", In: Proc. of the 12th ACM SIGKDD international conf. on Knowledge discovery and data mining, pp. 754-759. ACM, 2006.

[48]S. Xu,, J. Zhang, D. Han and J. Wang,. "Singular value decomposition based data distortion strategy for privacy protection", Knowledge and Information Systems, 10(3). pp. 383-397, 2006.

[49]X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud", Journal of Computer and System Sciences, 80(5). pp. 1008-1020, 2014.

[50]X. Zhang, LT. Yang, C. Liu and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud". IEEE Transactions on Parallel and Distributed Systems 25(2), pp. 363-373, 2014.

[51]J. Wang, W. Zhong, and J. Zhang. "NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets." Sixth IEEE International Conf. on Data Mining-Workshops (ICDMW'06), 2006.

[52]C. Wang, S. S. M. Chow, Q. Wang, K. Ren and W. Lou, "Privacy-preserving public auditing for secure cloud storage", IEEE Transactions on computers 62, no. 2, pp. 362-375,2013.

[53] SL. Warner, "Randomized response: A survey technique for eliminating evasive answer bias", Journal of the American Statistical Association 60.309, pp. 63-69, 1965.

[54] P. Zhou, J. Lei, and W. Ye. "Large-scale data sets clustering based on MapReduce and Hadoop", Journal of Computational Information Systems 7.16, pp. 5956-5963, 2011.