Privacy Preserving Data Mining Techniques: Challenges & Issues

Hitesh Chhinkaniwala¹ and Dr. Sanjay Garg²
¹Computer Engineering Department, U. V. Patel College of Engineering, Ganpat University, INDIA
²Computer Engineering Department, Nirma Institute of Technology, Nirma University, INDIA
E-mail: ¹hitesh.chhinkaniwala@ganpatuniversity.ac.in, ²sgarg@nirmauni.ac.in

Abstract—Privacy preserving becomes an important issue in the development of various data mining techniques. In this paper, we have discussed various techniques to preserve privacy while mining data. In the absence of uniform framework across all data mining techniques, researchers have focused on data technique specific privacy preserving issue. Available framework and algorithms provide further insight into future scope for more work in the field of fuzzy data set, mobility data set and for the development of uniform framework for various privacy preserving across all data mining algorithms.

Keywords: Privacy preserving; Cryptography; Randomization; data perturbation;

I. INTRODUCTION

Recent years have seen advancement in hardware technology and that leads to the increased capability to store and record personal data about consumers and individuals. Many business organizations collect data for analyzing business policy of competitor, consumer behavior and improving business strategies. Large amounts of such data can be used for different data mining purposes such as knowledge discovery, decision-making, statistical analysis etc. The potential societal benefits of data mining are substantial; however individual records are often considered to be private and sensitive. Steps must be taken to ensure that the privacy is not breached [31]. There are different perspectives of privacy to different people. Some may consider entire personal information as private while some may think certain attribute value should not be available directly or indirectly to the personal domain [40].

Privacy preserving data mining (PPDM) algorithms attempt to reduce the injuries to privacy caused by malicious parties during the rule mining process. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. All privacy-preserving transformations cause information loss, which must be minimized in order to maintain the ability to extract meaningful information from the published data. Privacy preserving data mining methods have been evaluated by research community and statistical disclosure committee mainly on the aspects like applicability, privacy protection metric, the accuracy of mining results, computation etc. [41]. Despite such recognition and efforts, there remain many challenges to foster data mining task while protecting individual privacy.

II. CLASSIFICATION OF PRIVACY PRESERVING DATA MINING TECHNIQUES

Various techniques have been proposed varied from entire data set modification or selective data set modification. K-anonymity offers selective data set modification to reduce granularity of the data using techniques such as generalization and suppression. Noise addition is another such approach where there is no dependency over underlying data and individual data from modified data set cannot be used for data mining purposes. Privacy preserving data mining techniques can be classified based on data distribution, data modification and privacy preserving techniques.

Data Distribution

- Centralized
- Distributed
  - Vertical Partitioning
  - Horizontal Partitioning

Data Modification

- Perturbation
- Aggregation
- Blocking
- Swapping
- Sampling
- Sanitization

Privacy Preserving Techniques

- Heuristic-based
- Reconstruction-based
- Cryptography-based

Fig. 1: Classification of Privacy Preserving Techniques
III. THE HEURISTIC-BASED APPROACH

Various techniques have been developed to sanitize or modify selective data for data mining techniques like association rule mining, classification and clustering. Selective data sanitization or modification based mining problem is NP-hard and for this reason, heuristic can be used to address the complexity issues. The concept of protecting respondent identity through micro data release using k-anonymity was first proposed by P. Samarati in [21], and subsequently many techniques have been proposed based on it, such as l-diversity [15], t-closeness [13], Incognito [12], and so on. k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. Homogeneity attack and back ground knowledge attack creates challenge to k-anonymity approaches. A dataset complies with k-anonymity protection if each individual’s record stored in the released dataset cannot be distinguished from at least k-1 individuals whose data also appear in the dataset. This protection guarantees that the probability of identifying an individual based on the released data in the dataset does not exceed 1/k. [52]

Samarati [21] first proposed k-attribution anonymity approach to protect respondents’ identities in micro data release. Author discussed how k-anonymity could be provided without compromising integrity of the information released using generalization and suppression techniques. How unknowns can be used in dataset to prevent discovery of sensitive association rules has been discussed in [22] with deterministic techniques only. Authors were concerned towards release of sensitive information and assuming that individual data privacy is already protected in released dataset. They proposed to use unknown instead of “False” value that produce misleading results. Microaggregation – Statistical disclosure techniques on quantitative micro data was introduced in [8]. Proposed disclosure techniques used heuristics and changed data while keeping statistical property intact. Instead of micro data, micro aggregates can be published. Information loss is major concern for such approach and it degrades the usability of data. This technique is limited to quantitative (Numeric) data and further modification is possible for categorical data. Chen X. et al. [6] proposed a framework based on dataset reconstruction using data sampling. Data sample is automatically generated using constraint-based inverse item set lattice mining technique. Authors addressed the problem of sensitive rule exposure and attempted to protect it by selective disclosure of frequent item set. The computation cost is mainly associated with construction and modification of item set lattice. The framework handles balance between privacy preserving and data sharing but face the problem of hiding failures, misses cost and artificial patterns (Ghost rules). Query auditing based privacy-preserving technique on statistical database introduced in [17]. Users considered one sensitive attributes and rest are publicly available. Auditor monitors queries by user (attacker) and denies if the answer compromise with privacy.

Bayardo R. and Agrawal R. first proposed an algorithm that finds optimal anonymizations under two representative cost measures and a wide range of attributes k [4]. Algorithm perturbs input dataset as little as is necessary to achieve k-anonymity. Authors also compared proposed algorithm with set of known approaches to the problem in terms of practicality and solution guarantees. Bayardo R. and Agrawal R. discussed optimal k-anonymization in [4] did not address the classification requirements. Friedman A. et al. [11] focused on k-anonymization for classification. This paper proposed only k-anonymization and it might not be optimal and is still an open issue.

Multidimensional k-anonymity model was proposed in [46] to preserve privacy in micro data publishing with more flexibility. Quality evaluation based on query workload explored in this paper. Greedy Approximation algorithm was used because Multidimensional k-anonymity is NP-hard problem. Authors concluded that greedy algorithm is more efficient than earlier proposed optimal k-anonymization algorithms for single dimensional model. Workload Driven Quality Measurement is a challenge where publish may want to consider anticipated workload.

k-anonymity model based approach failed to provide attribute level privacy. k-anonymity can create groups that leak information due to lack of diversity in the sensitive attribute. New definition of privacy called l-diversity was proposed by Machanavajjhala A. et al. [15] to overcome attribute level item set problem faced with k-anonymity. Proposed method seems to be biased towards privacy at the cost of usability. Entirely new approach was proposed in [13] which is an extension to known k-anonymity and l-diversity models. Authors proposed a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. Techniques such as K-anonymous, L-diverse and t-
closeness were given as solutions to solve the problem of privacy breach, at the cost of information loss [33]. Full-domain framework Incognito for popularly known k-anonymity model for multidimensional dataset proposed in [12] with set of algorithms to produce k-anonymous full-domain generalizations. Scalability of proposed framework- Incognito is still subject of concern.

Popularly known k-anonymity model was used by Fung C. M. et al. [10] and demonstrates how the definitions of k-anonymity can be applied to determine the anonymity of a decision tree. Proposed decision tree induction algorithms that are guaranteed to provide k-anonymity with better accuracy. Noise addition scheme in decision tree classification was explained in [27]. It also presented overview on decision tree and discuss evaluation metrics in terms of InfoGain and GainRatio. Different algorithms for noise addition into numeric attributes and categorical attributes were introduced. If Leaf Reaching Path attributes (LRPA) or Leaf Wrong Path Attributes (LWPA) are categorical then CAPT (Categorical Attribute Perturbation Technique) has been used. On the other hand two separate algorithms for noise addition have been proposed for numerical attributes based on whether they are part of LRPA or LWPA. Data quality in modified data is still a big question with Noise addition techniques and proposed set of algorithms work only for decision tree based classification techniques.

Zhong S. et al. [25] used k-anonymization of distributed customer data with cryptography techniques. Proposed protocol claims to provide end-to-end privacy. Blum A., et al discussed SuLQ framework for query auditing in [5], where trusted administrator prevents access to private information by adding noise to the query response. This approach does not perturb input data at all. Junwei Z. et al. [34] proposed framework for privacy mining using k-anonymization data stream. Sliding window mechanism has been used to collect data that fall within. Framework use distribution density to improve the precision of data and in turn increases the usability. Poovannam E. et al. [44] focused with medical dataset as being heterogeneous in nature and it has great human, social ethical and legal values. Protection of certain sensitive attributes while mining is of prime concern. Detail analysis with possible attacks has been elaborated by authors on k-anonymity, i-diversity and t-closeness approaches. Unlike k-anonymity proposed work focus on sensitive data protection only. Methods can be applied to both numerical and categorical data types. Correlation has

been measured between original data and transformed data. As correlation factor is very high, any researcher having the rules/results on transformed dataset can easily get the actual values from central server.

Lot many approaches have been popped up for association rule hiding by decreasing the main factors of association rule mining i.e. support and confidence [1][7][18][19][24][39]. Knowledge hiding issue for the first time in sequential pattern mining was introduced in [1] and proved that finding an optimal sanitization of dataset is NP-hard. Followed heuristic approach (Greedy search) to hide individual sensitive rules instead of all rules produced by some sensitive item sets. Better heuristic approach can be possible to improve upon the performance.

Authors introduced framework to balance trade-off between privacy and accuracy in [19]. Rather than adding noise, authors selectively remove certain individual items so as to restrict sensitive rules to appear as an output. Applicability of proposed approach to other data mining techniques - classification and clustering is still an open issue. There is no guarantee of non-sensitive association rules generation after applying algorithms. Algorithms are also generating artificial patterns (ghost rules). In another work, Oliveira S. and Zaiane O. [18] used data sanitization techniques to preserve individual data privacy from association rule mining techniques. One scan sliding window algorithm was proposed and has been more efficient compare to one proposed by [7] on hiding association rules. Authors claimed that Sliding Window algorithm has lowest misses cost among peer algorithms. Duraiswamy K. [24] proposed an algorithm (SRH) to hide the sensitive rules that contain sensitive items, so that sensitive rules containing specified sensitive items on the right hand side of rule cannot be inferred through association rule mining. First step generate association rules using AIS algorithms and clusters the rule using method proposed in [18]. There is vast scope of efficiency improvement and still algorithm has side effects. Manoj G. and R. C. Joshi presented method for fuzzy rules hiding in quantitative data in [39]. Fuzzy association rule hiding has been less explored area and techniques for privacy protection for fuzzy rule hiding is still emerging. Authors have focused on association rule mining techniques only. Single membership function has been used to compute the fuzzy value from input dataset. The algorithm tries to decrease the support of A $\rightarrow$ B by decreasing the frequency of item set (AB) until either support or confidence value goes below minsup or minconf respectively. Further the support of item set (AB) can be decreased by decreasing the support count of either A or B. The
performance of the algorithm has been measured according to three criteria: Number of rules hidden, database effects (how the modified fuzzy value affect usability of dataset and in turn hide non-sensitive rule along with sensitive rules) and side effect produced (ghost rule generated). Approach can be expanded with membership function for individual attributes from dataset.

IV. THE RECONSTRUCTION-BASED APPROACH

Reconstruction-based techniques perturb the original data to achieve privacy preserving. The perturbed data would meet the two conditions. First, an attacker cannot discover the real original data from the issuance of the distortion data. Second, the distorted data is still to maintain some statistical properties of the original data, namely some of the information derived from the distorted data are equivalent to data obtained from the original information [42]. So for each data mining techniques separate algorithms need to be developed. Data perturbation is that the value of each property is transformed into other value of the property domain by given probability [41]. The drawback of such methods is that they impair the data integrity, and this drawback is well studied in k-anonymity methods in various research papers.

Agrawal and Srikant first proposed [3] reconstruction-based technique for privacy of sensitive values. Sensitive values were perturbed using randomization function so that original values cannot be estimated with sufficient precision. But authors did not address the effectiveness of randomization with reconstruction for categorical attributes. In another approach for hiding sensitive rules, Rizvi and Haritsa [20] presented a scheme called MASK (Mining Association with Secrecy Constraints) that distorts original data with measured probability before it is available to the data miner. In other work [9], random data distortion (perturbation) technique was used to prevent disclosure of sensitive personal information at the same time making sure that the underlying distribution is accurately preserved.

This approach increased the usability of perturbed data if relative amount of noise is added to actual data but under certain conditions it is relatively easy to breach the privacy protection. Privacy preserving data mining has been discussed for specialized application where users do no computation and data flow is only from the users to the data mining agent. The same line of conclusion was made in [38], elaborating privacy lose and information loss properties in data perturbation scheme. A random perturbation based framework (FRAPP) keeping in mind the tradeoff between privacy and accuracy was proposed by Agrawal and Haritsa in[2]. Investigation shows that the tradeoff is very attractive in that the privacy increase is substantial whereas the privacy reduction is only marginal also demonstrated the utility of proposed schemes in the context of association rule mining. The systematic identification of perturbation matrices and the conditions on their applicability is still an open research issue. Different approach of Privacy via pseudo random sketches was discussed in [16]. Research work summarized that error of approximation is independent of the number of attributes involved and only depends on the number of users available. User releases data in sketches -- subset of user’s attributes. Authors limit their analysis to only AND queries because it is very powerful to extract private information using linkage.

All methods discussed so far released distortion parameters for statistical analysis task but disclosing distortion may breach privacy. Guo and Wu [30] used randomized techniques for privacy preserving with unknown distortion parameters. Their approach was carried out on categorical data only and extension of it on numerical and network data is still topic of research. A new scheme was introduced in [28] which have used both randomization and data perturbation techniques. Major attraction of the scheme is its individual adaptability in choosing the privacy levels for different attributes during perturbation. Experimental study shows that scheme is biased towards privacy and still further work is needed to apply approach on distributed data. Gaussian distribution technique to preserve privacy before data classification was proposed in [45]. According to this approach random value has been generated using Gaussian approach, noise will then be added into generated variables to make it impossible for anyone to regenerate original value from perturbed dataset. Classification tree has been generated on perturbed dataset upon receiving user query. Proposes technique applied to numerical data only and there is much scope of improvement in suggested approach.

Xiaolin Z. et al. [41] presented an approach for privacy preserving classification data mining based on the matrix of random perturbation, which is applicable to any data distribution. R-amplifying method and matrix condition number have been used to protect data privacy. Proposed method can be applicable to all kind of data types and improved privacy protection and accuracy of mining results.
V. THE CRYPTOGRAPHY-BASED APPROACH

In a distributed environment the primary issue to achieve privacy preserving is the security of communication, and encryption technology just to meet this demand. Therefore, privacy preserving based on data encryption technology commonly applies to distributed applications. Lindell & Pinkas [14] first proposed Secure Multi-party Computation protocol for data mining classification techniques. Cryptography-based techniques offer a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exist vast toolset to perturb dataset [43]. Cryptography-based techniques have more time complexity compare to other methods for data updating. Further on application front, Lindell [29] demonstrated that how privacy preserving data mining can be incorporated with smart cards. Cryptography techniques are used to preserve privacy.

Lindell & Pinkas [14] first discussed ID3 algorithm and Decision Tree classification with privacy preservation. Research work was carried out with secure two-party computation protocol extending ID3. Communication overhead was considered as big drawback for cryptography-based privacy preserving techniques hence there is always scope of improvement. Vaidya and Clifton [23] came up with partitioning approach for data residing at different location and are used for data mining. They proposed a method of partitioning data vertically and store them on various sites. This way each site learns the cluster of each site. This way each site learns the cluster of each entity, but learns nothing about the attributes at other sites. Paper discussed a method for k-means clustering and demonstrates how results from Secure Multi-party Computation (SMC) can be used to generate privacy-preserving data mining algorithms. Proposed approach still face challenge for possible tradeoff between communication cost and level of privacy. Further work on protocol suggested by Lindell & Pinkas for collaborative data mining was carried out in [26,35]. Framework systematically transforms normal data mining computation to Secure Multi Party Computation (SMC) introduced by Lindell & Pinkas in [14]. Mapping table and Graded grouping techniques were used to transform numeric and categorical data.

Enhancement and similarity measure for two encrypted data points under semi-honest environment was proposed in [36], Jaccard similarity function and Private Equality Test (PET) have introduced with reasonable complexity. Authors claim to have reduction in communication cost compare to other cryptography-based techniques. Paper focused only on similarity measure using Jaccard notation and result validating through implementation is still pending. XueFling et al. [37] introduced parallel algorithm using combined heuristic-based and cryptography-based privacy protection techniques. Proposed techniques can realize the concealment of frequent item set and in turn protect the privacy of association rule mining.

VI. CONCLUSION

Existing algorithms offering privacy to the extraction of knowledge patterns and need further investigation for possible improvements. Common framework is still an issue that will unify move advanced measures for the evaluation and the comparison of different privacy preserving data mining methodologies. Privacy preserving data mining for fuzzy item set values still has more open issues to gain privacy at the same time reduction in information loss. Mobility data mining and privacy-aware stream data mining are among the most recent and prominent directions of privacy preserving data mining. Privacy in the context of applications where data release incrementally and in an unconditional rate is creating major challenges to the data mining community.

REFERENCES

Privacy Preserving Data Mining Techniques: Challenges & Issues

30. Guo L. and Wu X., Privacy preserving categorical data analysis with unknown distortion parameters, Transactions on Data Privacy 2, pp. 185–205, 2009.