Privacy Preservation Techniques in Data Mining: A Survey

[1]Aanchal Sharma, [2]Neha Rani, [3]Dr. Sudhir Pathak

Apex Institute of Technology
Chandigarh University, Mohali

Abstract: **The privacy preservation is an essential process of saving information during the data mining and it is rising as the technology is growing up. This approach is used to protect the sensitive information and provides the security to the data. Different types of approaches like randomization, encryption and anonymization have been proposed for data protection and information mining. This review paper presented the different privacy preserving approaches used by the researchers in collaborative data mining. The parameters analysis of the different approaches is also given in this work.**

IndexTerms: **privacy, preserving, precision, datamining**

I. INTRODUCTION"

Now the day, due to advancement in technology it is easy to access the data information from anywhere which is store on the server. These types of technology reduce the storage cost, time and effort. There are many data mining techniques that are used for analysis of data which increased in a huge amount. Basically data mining techniques is used to extract the information from the huge amount of data and preserve the private details of that data.

Privacy preservation is the most leading research field in the field of data security in cloud. A number of algorithms and techniques are introduced for preserve the privacy of data. It is essential to maintain the ratio between privacy protection and knowledge discovery. The most common privacy preserving techniques are randomization method, encryption method and anonymization method. Random method masks the values of the records by adding the mask data to the original data. Encryption method changes the data into the meaningless form and protect from the attackers. In anonymization it makes record indistinguishable among the group records by using generalization approaches.

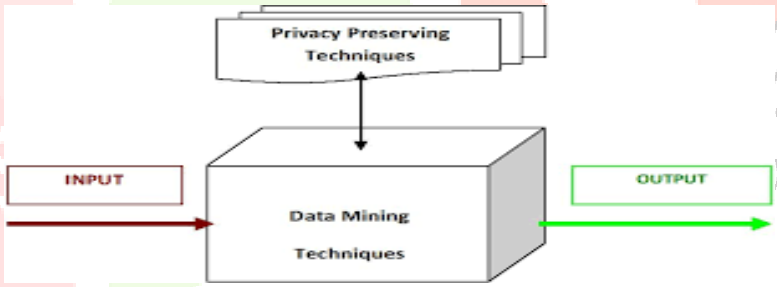


Figure 1.1 Privacy preservation

I. LITERATURE SURVEY

Data mining methods are analyzed on the medical data context. In this process data mining depend on the Euclidean distance and patterns. In this kernel methods and spectral clustering is also used for mining data with privacy. The proposed work solves the problem of decentralized matrix problem by using different methods like random multiplicative updates which provides the secure computation of distances [1]. Privacy Preserving data mining techniques are used to preserve the privacy of the large amount of data. This method of privacy preserving is helpful when different parties want to done collaborative data mining. In this work author reviewed the different approaches used for the privacy preservation [2]. Non-negative matrix factorization and singular value decomposition approach is proposed for data protection with privacy. These methods analyze data and find the more critical information. Data quality is also maintained under this work with privacy preservation [3]. Privacy preserving algorithm is proposed which is based on frequent path for trajectory data. TOPF method solves the problem of privacy preservation and provided effective data usability. In this work a new method is proposed for most frequent path finding and divide trajectory into candidate groups [4].

The K-mean clustering algorithm is proposed which solves the problem of clustering in the data mining field. The comparative study of K-mean, cobweb and hierarchical is performed on two datasets and clusters are formed by using WEKA API. Privacy of data is maintained by using encryption and decryption method with AES algorithm. Modified K-mean method performs better than K-mean method. Modified K-mean method forms effective number of clusters in organized manner and takes less amount of time in the whole process [5]. In this work, a general framework is proposed for privacy preserving sequential data. The proposed algorithm handles the data and protects this data from linking attack. It maintains the i-diversity and increases the data utility during data publication. This method provides the effective data utility the existing method [6].

Distributed Deep Learning approach is introduced to preserve the privacy in the smartphones. Learning approach is also used from preserving the privacy. The machine learning techniques are mostly based on cloud and their services which also enhance the risk of the attack. To protect the data from these types of attack hybrid deep learning method is effectively worked on it [7]. Double projection deep computation model is proposed for big data feature learning with algorithm. The efficiency of the training model is enhanced by using learning algorithm and deep computation method preserves the privacy using BGV encryption method [8].

The Rampart's framework is used to categorize protection approach and find out the effective from them for data mining. In this author worked on the knowledge discovery of data (KDD) and extract the patterns. Rampart's frame work also generates cluster and Support vector machine is used for data classification process [9]. Anonymization based approach is introduced for the privacy preserving of the information. In this work greedy k-member algorithm and systematic clustering algorithm provides the less information loss. These algorithms minimize the information loss risk by using systematic clustering algorithm. The effectiveness of the proposed method is evaluated by comparing with existing methods [10].

Table 1: Inference from the literature review

Author's Name	Year	Technology/ Algorithm Used	Observations.
1. Scardapane, Simone, et al.	2018	Kernel Methods and Spectral Clustering	The proposed work solves the problem of decentralized matrix problem by using different methods like random multiplicative updates which provides the secure computation of distances [1]
2. Toshniwal et al.	2018	Privacy Preserving Data Techniques	In this work author reviewed the different approaches used for the privacy preservation. The proposed work solves the problem of decentralized matrix problem by using different methods like random multiplicative.
3. Li, Guang et al.	2018	Matrix Factorization and Single Value decomposition	These methods analyze data and find the more critical information. Data quality is also maintained under this work with privacy preservation.
4. Dong et al.	2018	Privacy preserving algorithm based on frequent path.	TOPF method solves the problem of privacy preservation and provided effective data usability. In this work a new method is proposed for most frequent path finding and divide trajectory into candidate groups
5. Khan, Shifa et al.	2018	K-mean Approach	The comparative study of K-mean, cobweb and hierarchical is performed on two datasets and clusters are formed by using WEKA API. Modified K-mean method forms effective number of clusters in organized manner and takes less amount of time in the whole process.
6. Hasan et al.	2017	Sequential data publishing	The proposed algorithm handles the data and protects this data from linking attack. It maintains the i-diversity and increases the data utility during data publication. This method provides the effective data utility the existing method
7. Ossia, Seyed Ali, et al.	2017	Deep Learning Approach	Distributed Deep Learning approach is introduced to preserve the privacy in the smartphones. Learning approach is also used from preserving the privacy.

8. Zhang, Qingchen, et al.	2017	Deep Computation Model	Double projection deep computation model is proposed for big data feature learning with algorithm. The efficiency of the training model is enhanced by using learning algorithm and deep computation method preserves the privacy using BGV encryption method.
9. Xu, Lei, et al.	2016	Support vector machine	In this author worked on the knowledge discovery of data (KDD) and extract the patterns. Rampart's frame work also generates cluster and Support vector machine is used for data classification process.
10. Bhaladhare et al.	2016	K-anonymity Model	Anonymization based approach is introduced for the privacy preserving of the information. In this work greedy k-member algorithm and systematic clustering algorithm provides the less information loss.
11. Zhang, Xuyun, et al.	2015	Anonymization Approach	Local recoding problem is solved by using the annonymization approach and two phase clustering approach is used to make the clusters. Scalability is enhanced by using Map reduce approach.
12. Prakash et al.	2015	Anonymization Approach	It protects the data from skewness, homogeneity, similarity and background attacks. Used Annoymzation model to preserve the privacy of the data.
13. Yang, et al	2015	Hybrid privacy preserving approach	A privacy preservation model is proposed for the medical data in cloud computing. It performs four basic function that are vertical data partitioning, data merging, integrity checking and the hybrid search performed on plan text and cipher text.
14. Sun, Li, et al.	2015	Support Vector machine and Vertical Partitioning	The proposed approach gives better security results by using SVM kernels and provides effective accuracy and classification. It reduces the computational time of the whole process.
15. Yun, Unil, et al.	2015	Fast Perturbation Algorithm	It uses FP algorithm for privacy preservation. The search is based on the tree structure. It reduces the search space of PPUM by using two traversal tables. This algorithm gives five to ten time faster result then the existing approach.

Table 2: Methods in the related work

	Cryptography	Fuzzy logics	Neural Network
Random Perturbation[2]	Yes	Yes	NO
K-anonymity[5]	Yes	Yes	Yes
Horizontally portioned Distribution[7]	Yes	Yes	Yes
Clustering[9]	Yes	Yes	Yes
Classification[10]	Yes	Yes	Yes
Association Rule mining[12]	Yes	Yes	Yes
Secured Sum Computation[13]	Yes	Yes	Yes
Aggregation[15	Yes	Yes	NO

Table 3: Performance parameters in related works.

Research Work	Parameters			
	Accuracy	Precision	Recall	F-Measure
[1]	×	×	×	√
[2]	√	×	×	×
[3]	√	×	×	×
[4]	×	×	×	√
[5]	√	×	×	×
[6]	×	×	√	×
[7]	×	√	×	×
[8]	√	×	×	×
[9]	×	×	×	×
[10]	√	×	×	√
[11]	√	√	√	×
[12]	×	×	×	√
[13]	×	×	×	×
[14]	×	√	√	×
[15]	√	√	×	×

Privacy Preserving Framework

The main issue that arises these days is privacy preservation of data because when collaborative data mining is performed some sensitive data is lost. The owner of the data also wants to hide some sensitive data while using data mining process. Privacy preservation in data mining is a process of mining useful data and also preserve the privacy of the sensitive data

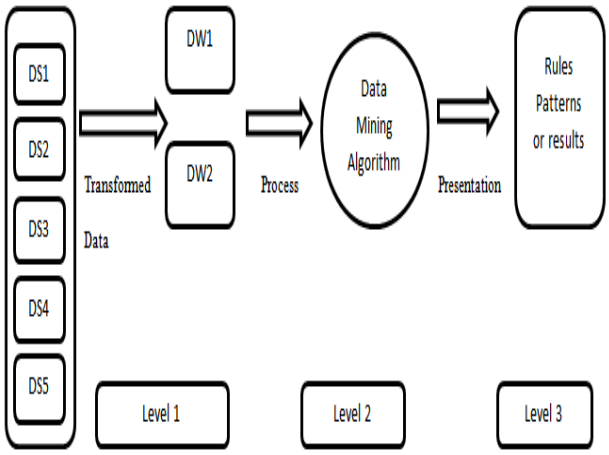


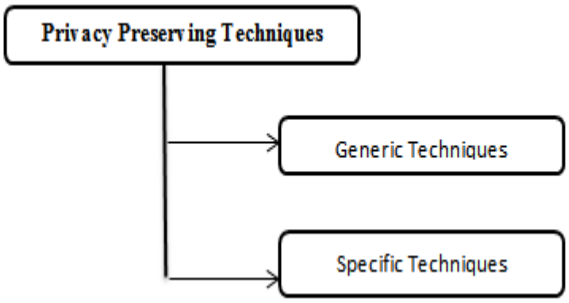
Figure 1.2 Framework of Privacy Preserving.

In privacy preserving framework firstly data is collected from the various sources and stored in the data warehouse. After this convert this data into suitable format for analytical process and then apply the data mining techniques on data. In first level data is collected from the various sources and then tested it is suitable for further processing or not. In level 2, Data sanitization is

performed for the privacy preservation. In sanitization process sampling, blocking, perturbation and generalization is performed on the data and apply the data mining algorithm. In level 3, data information is checked in order to maintain the privacy.

II. PRIVACY PRESERVING TECHNIQUES

There are two types of the privacy preserving techniques that are following [3].



- **Generic Techniques**: These are the approaches which introduced the privacy preservation into data in such a way that the transformed data can be used as input to perform any data mining task. Example of generic techniques are generalization, discretization etc.

- **Specific Techniques**: These techniques used algorithm for which performs the data mining and hiding the sensitive data. Some techniques of specific types are like KNN, secure multiparty computation etc.

III. CONCLUSION

Privacy Preserving data mining techniques are used to preserve the privacy of the large amount of data. This method of privacy preserving is helpful when different parties want to done collaborative data mining. In this work author reviewed the different approaches used for the privacy preservation In this paper a wide review on different privacy preserving approaches like fuzzy logic, neural network, particle swarm optimization etc. The table in the paper shows the effect on the parameters by the approach used. This review likewise causes analysts to comprehend the fundamental parts played by Fuzzy rationale, neural system, Cryptography and secure aggregate calculation technique.

REFERENCES

[1] Scardapane, Simone, et al. "Privacy-preserving data mining for distributed medical scenarios." *Multidisciplinary Approaches to Neural Computing*. Springer, Cham, 2018. 119-128.

[2] Toshniwal, Durga. "Privacy Preserving Data Mining Techniques for Hiding Sensitive Data: A Step Towards Open Data." *Data Science Landscape*. Springer, Singapore, 2018. 205-212.

[3] Li, Guang, and Rui Xue. "A New Privacy-Preserving Data Mining Method Using Non-negative Matrix Factorization and Singular Value Decomposition." *Wireless Personal Communications* (2018): 1-10.

[4] Dong, Yulan, and Dechang Pi. "Novel Privacy-preserving Algorithm Based on Frequent Path for Trajectory Data Publishing." *Knowledge-Based Systems* (2018).

[5] Khan, Shifa, and Deepak Dembla. "Implementation of Modified K-means Approach for Privacy Preserving in Data Mining." *Advances in Computer and Computational Sciences*. Springer, Singapore, 2018. 601-610.

[6] Hasan, p;ASM Touhidul, and Qingshan Jiang. "A general framework for privacy preserving sequential data publishing." *Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on*. IEEE, 2017.

[7] Ossia, Seyed Ali, et al. "A hybrid deep learning architecture for privacy-preserving mobile analytics." *arXiv preprint arXiv:1703.02952* (2017).

[8] Zhang, Qingchen, et al. "Privacy-preserving double-projection deep computation model with crowdsourcing on cloud for big data feature learning." *IEEE Internet of Things Journal* (2017).

[9] Xu, Lei, et al. "A framework for categorizing and applying privacy-preservation techniques in big data mining." *Computer*49.2 (2016): 54-62.

[10] Bhaladhare, Pawan R., and Devesh C. Jinwala. "Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model." *J. Inf. Sci. Eng.* 32.1 (2016): 63-78.

[11] Zhang, Xuyun, et al. "Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud." *IEEE transactions on computers* 64.8 (2015): 2293-2307.

[12] Prakash, M., and G. Singaravel. "An approach for prevention of privacy breach and information leakage in sensitive data mining." *Computers & Electrical Engineering* 45 (2015): 134-140.

[13] Yang, Ji-Jiang, Jian-Qiang Li, and Yu Niu. "A hybrid solution for privacy preserving medical data sharing in the cloud environment." *Future Generation Computer Systems* 43 (2015): 74-86.

[14] Sun, Li, et al. "A new privacy-preserving proximal support vector machine for classification of vertically partitioned data." *International Journal of Machine Learning and Cybernetics* 6.1 (2015): 109-118.

[15] Yun, Unil, and Jiwon Kim. "A fast perturbation algorithm using tree structure for privacy preserving utility mining." *Expert Systems with Applications* 42.3 (2015): 1149-1165.