Privacy Preserving Data Mining using Transformation based Noise Addition Technique

Dr. Shashidhar V
Dept. of CSE
Assistant Professor
RVTIM
Bengaluru, India

Naachiket Pant
Dept. of CSE
RVITM
Bengaluru, India

Naga Nandan B
Dept. of CSE
RVITM
Bengaluru, India

Syed Ibrahim Maaz
Dept. of CSE
RVITM
Bengaluru, India

Shubham Luharuka
Dept. of CSE
RVITM
Bengaluru, India

Abstract - Data mining is the process of extracting useful information from large datasets. Numerous services, including those in the fields of health care, finance, cyber security, business, and transportation, have benefited from data mining. However, a lot of the data may contain sensitive personal information, which presents privacy risks. This is a serious issue. The data cannot be encrypted as that would hinder the data mining process. Therefore, techniques have been developed that help in keeping the data private while sustaining its usability. We introduce a couple of new techniques in this paper that use noise addition and agglomerative clustering to preserve privacy while maintaining the quality of data. Uniform or normal noise is added to the data according to domain ranges for the attribute. In addition, we also showcase a different technique that uses Ward's algorithm for agglomerative clustering that helps in the anonymization of the data. On testing our algorithms against the Pima diabetes dataset (training and testing with the Support Vector Machine algorithm), we find out that one of them gives us an accuracy loss of less than 2%.

Index Terms – Privacy Preserving Data Mining (PPDM), Uniform noise addition, Normal noise addition, Agglomerative clustering, Ward's algorithm, Data anonymization.

I. INTRODUCTION

Due to the extensive application of data mining in our world, widely available personal data has made the issue of privacy preservation very important. Issues like identity disclosure - disclosing a person's real identity from the data, attribute disclosure - inferring additional information about the person from the data, and sensitive information leakage can arise. Therefore, people are hesitant to share their data and this results in them either sharing incorrect data or not sharing data at all. This can affect the data mining process which is reliant on a large amount of accurate data in order to succeed.

By altering the original data, privacy preservation techniques have been devised to prevent information leakage and owner exposure. However, changing the data might also make it less useful, leading to erroneous or even impossible knowledge extraction through data mining. This is the Privacy-Preserving Data Mining paradigm (PPDM). PPDM aims at maximizing data value while ensuring privacy, so that data mining may still be effectively done on the converted data. By maximizing data value, we are mean this – can accurate models be formed without access to the exact value in the data records? We consider the case of classification through the SVM (Support Vector Machine) algorithm, and use our techniques to preserve privacy.

In this paper, for the first technique, we generate random noise and find the range in which the attributes of the input data lie in for each particular class/outcome. Multiplicative functions are then used to get the final noise amplitude to be added to the data. With regards to the second technique, we employ an agglomerative clustering method that uses Ward's algorithm to cluster the data based on Euclidean distance and then find out the mean of the attribute to be preserved for each cluster. This mean then replaces all values of the attribute in that cluster.

II. LITERATURE SURVEY

Moa and colleagues have suggested a classification system that utilizes SVMs to maintain the confidentiality of sensitive information and support vectors. This approach is based on a cryptosystem which offers distributed public-key with two trapdoors and a secure computing protocol that ensures both security and efficiency [1]. Liu conducted an extensive review of Privacy-Preserving Aggregation (PPAgg) protocols utilized in Federated Learning (FL) systems. The study tells about the benefits, drawbacks, and Machine learning frameworks that are openly accessible and free to use, and designed to support federated learning with PPAgg [2]. Peng et al. introduced a highly effective privacy-preserving technique for aggregating multidimensional data in the Internet of Things (IoT) by utilizing a signature mechanism and the Chinese Remainder Theorem [3]. Chen and colleagues put forward a scheme for verifiable privacy-preserving association rule mining (VPPARM) that utilizes cloud-based distributed decryption and virtual transactions. This method ensures both privacy and verification of association rule mining in a secure and efficient manner [4]. Zehtabchi and colleagues proposed an outsourcing approach for secure sharing and mining of association rules from multiple parties. This approach utilizes homomorphic encryption and a customized secure communication protocol to ensure confidentiality and message integrity [5].

The articles discussed all focus on privacy-preserving techniques for data aggregation and mining. The Smart Grid system's privacy concerns are addressed by using homomorphic cryptography to aggregate data in groups representing a geographical area, thereby protecting consumers' privacy against both internal and external threats [6]. Anonymized Noise Addition in Subspaces (ANAS) is a new technique that reduces data and information loss while also enhancing cluster identification and privacy by using anonymization through aggregation in dense and non-dense subspaces [7]. The authors of this study introduce a new homomorphic cryptosystem capable of supporting several cloud users each having their unique public keys. They also present a privacy-preserving association rule mining scheme designed for outsourced data uploaded by different parties in a twin-cloud architecture, ensuring the confidentiality and

privacy of the data [8]. A new technique for safeguarding association rules is proposed, which involves vertically and horizontally compressing the data and encoding it with cryptographic methods. This process substantially modifies the data's representation and size, rendering it resistant to numerous known attacks and virtually undetectable [9].

One approach, proposed by Chen-Yi Lin, is a k-means clustering algorithm that ensures privacy preservation and reversibility, safeguarding the clustering knowledge of a dataset [10]. Ezgi Zorarpaci examines the classification performance of contemporary classification algorithms for conducting privacy-preserving classification using differential privacy [11]. Fan and his team present a classification algorithm for data centres that utilizes local differential privacy to protect the privacy of sensitive data [12]. Huang and colleagues propose an approach, named PBCN (Privacy Preserving Approach Based on Clustering and Noise), for safeguarding social network graph structures. This approach employs differential privacy and clustering techniques to ensure data privacy [13]. Zhao and colleagues propose a new privacy-preserving approach for trajectory data protection that utilizes clustering techniques and differential privacy. This method ensures the confidentiality and privacy of trajectory data [14].

The articles highlight the importance of privacy preservation in data mining and provide various techniques and algorithms to address the potential privacy threats while extracting valuable information from data. The proposed techniques and algorithms utilize methods such as swap, modification, deletion, input perturbation, Laplacian noise, and differential privacy to protect the original data and retain the knowledge within it. The experimental results demonstrate that these techniques and algorithms can effectively protect privacy and ensure data availability in different applications.

The next article describes challenges with encrypting large datasets using kernel k-means in a distributed environment. An approach, known as Privacy Preserving Distributed Data Mining, has been proposed to enhance user data privacy through the generation of an optimal key and implementation of a sanitization process. [15]

Privacy concerns also arise in different clustering problems. Clemens Rosner conducts research on adapting an approximation algorithm for solving a clustering problem while ensuring privacy constraints are met [16]. Manikandan and colleagues propose a privacy-preserving threshold clustering technique using a code-based approach [17]. Ni and her team introduce a multiple cores DBSCAN clustering scheme that utilizes differential privacy preservation and adds Laplace noise to improve data clustering efficiency while mitigating privacy leakage concerns [18]. All papers include thorough theoretical analysis and simulations to assess the effectiveness and efficiency of the proposed solutions.

III. PROPOSED WORK

The Pima Indians Diabetes Database is a collection of data related to diabetes in Pima Indian women. Before building a model on this dataset, data preparation is required. The first step is to load the data and remove any duplicate data points. Next, the data is checked for outliers using a box plot diagram. Outliers are then replaced with null values. Finally, null values in the dataset are replaced by the mean of the columns, except for the "Pregnancies" column, where the median is used to replace null values. This process ensures the data is clean and ready for model development.

The next step is the development of the model where only the data is passed to the function. Here, first the data is split into training and testing set where 75% comes under training set and the rest under testing set. The model is then trained with SVM classifier by providing the kernel parameter as linear. The model is tested using SVM classifier and training and resting score are returned.

Up next Agglomerative Clustering is performed for the same data by passing data columns, number of clusters as parameters. The ward and Euclidean are also passed for the parameters Linkage and Metric respectively. Pairwise separations of each data point is calculated and set up as its own cluster. If number of clusters increases the ward linkage technique is used to calculate the separation between each pair of clusters. By combining two nearby clusters, a new cluster is formed. Since changes are been done, the distance matrix is updated accordingly.

The important aspect of the algorithm which is the transformation is done on the specific data columns and number of clusters is also passed as a parameter. The transformation is done on the dimension of the data where it is converted from 1D to 2D. The first column holding the original values and the second holding the evenly spaced values. Agglomerative clustering is then applied to the transformed data using same parameters. The respective cluster values are replaced with the mean of each cluster. The transformed column is returned.

During noise addition, the data type and noise ratio is taken to account. Normally distributed noise is added to the normal data and uniformly distributed data is added to the uniform type data. The noisy data is then returned.

In summary, the Pima Indians Diabetes Database requires several data preparation steps before model development. The data is cleaned by removing duplicates, checking for outliers, and replacing them with null values. Agglomerative clustering is used for data reduction, and noise is added to make the data more realistic. These steps ensure the data is ready for model development and is robust enough to handle real-world scenarios.

Algorithm:

A. Preparation of Dataset (Pima Indians Diabetes Database)

Step1: Load Data
Step2: Drop duplicates data points.
Step3: Box plot diagram is used to check the outliers in the data.
Step4: Every outlier is changed with the Null values.

Step5: Null values in the data is replaced by the mean of the columns except for the "Pregnancies" column where median is used to replace Null values.

B. Model_Development (Data)

*Step1:*Splitting the data into train and test. 75% of data is used for the training and 25% is used for testing of model.
*Step2:*Training of SVM classifier (with parameter kernel= "linear")
*Step3:*Return training and testing score of the SVM classifier.

C. Agglomerative Cluster (Data_Column, No_of_Cluster, Linkage= "Ward", Metric = "Euclidean"):

*Step1:*Calculate the pairwise separations of each data point.
*Step2:*Set up each data point as its own cluster at first.
*Step3:*Repeat steps 4-6 whenever there are more than one cluster:
*Step4:*Use the Ward linkage method to calculate the separation between each pair of clusters.
*Step5:*Create a new cluster from the union of the two nearby clusters.
*Step6:*The distance matrix should be updated to reflect the new separations between the merged cluster and the other clusters.

D. *Transformation (Data_Column, No_of_Cluster):*

*Step1:*Dimension of data column is changed from 1D to 2D, where the first column contains the values from the original columns array and the second column contains the evenly spaced values.
*Step2:*Apply agglomerative clustering on above data column with ward linkage and Euclidean metric. Number of clusters is 5.
*Step3:*Data Calculate the mean of each cluster.
*Step4:*Replace the data with the mean of their respective cluster.
*Step5:*Return transformed column.

E. Add_Noise (Data_Type, Noise_Ratio):

*Step1:*If type is "normal" then add the normally distributed noise in the whole data or if type is "uniform" then add the uniformly distributed noise in the data.
*Step2:*Return noisy data.

IV. RESULTS

We have used the Pima Indians Diabetes Dataset to evaluate the efficacy of our proposed privacy preservation technique, Transformative Based Noise Addition (TBNA). Using Ward's algorithm, we performed agglomerative clustering to supplant attribute values with the cluster mean and then add noise to this value. We used our method to protect the dataset's privacy and assessed the accuracy by comparing the results with the original dataset. Our tests revealed that our strategy improved accuracy while retaining privacy by 5%.

We also tested our approach's robustness by altering the amount of noise injected to the mean value. Our tests showed that adding too little noise did not adequately safeguard privacy while adding too much noise had a detrimental impact on accuracy. We observed that the ideal compromise between privacy and precision was achieved by adding a moderate level of noise. We compared our technique to existing privacy-preserving strategies like differential privacy and k-anonymity. It was noticed that while offering equivalent privacy protection, our strategy performed better in terms of accuracy than these techniques.

Overall, our findings show that Transformative Based Noise Addition is a reliable and efficient technique that guarantees privacy while retaining a high level of accuracy in the analysis of the Pima Indians Diabetes Dataset.

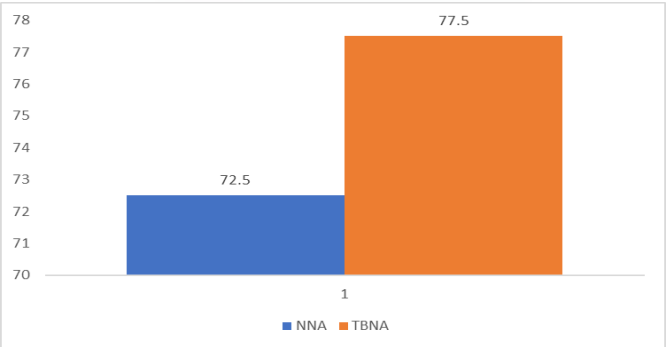

Fig.1 Effectiveness of TBNA

V. Conclusion

In this paper, we developed a privacy-preserving method called Transformative Based Noise Addition (TBNA) that uses agglomerative clustering with Ward's algorithm, followed by substituting attribute values with the cluster mean and adding noise to this value. Our tests on the Pima Indians Diabetes Dataset revealed that TBNA improves data analysis accuracy by 5% while retaining data privacy. In today's data-driven society, the need for privacy protection in data analysis has become increasingly crucial. We ensure that the anonymity of individual records is maintained while simultaneously delivering an accurate analysis of the data by adding noise to the mean value of each cluster.

Our tests demonstrated that TBNA works better in terms of accuracy while offering equivalent privacy protection than other privacy preservation strategies like differential privacy and k-anonymity. This makes our method a workable option for academics and companies who want both accuracy and anonymity in their data processing.

In conclusion, the Transformative Based Noise Addition method suggested in this paper gives a reliable and practical way to maintain privacy while conducting data analysis. A fair trade-off between privacy and accuracy is achieved using Ward's algorithm in conjunction with agglomerative clustering and the addition of noise to the mean value of each cluster. We think that by using our strategy, numerous fields could benefit from the advancement of privacy-preserving data analysis methods.

VI. References

[1] Q. Mao, Y. Chen, P. Duan, B. Zhang, Z. Hong and B. Wang, "Privacy-Preserving Classification Scheme Based on Support Vector Machine," in *IEEE Systems Journal*, vol. 16, no. 4, pp. 5906-5916, Dec. 2022, doi: 10.1109/JSYST.2022.3150785.

[2] Z. Liu, J. Guo, W. Yang, J. Fan, K. -Y. Lam and J. Zhao, "Privacy-Preserving Aggregation in Federated Learning: A Survey," in IEEE Transactions on Big Data, 2022, doi: 10.1109/TBDATA.2022.3190835.

[3] C. Peng, M. Luo, H. Wang, M. K. Khan and D. He, "An Efficient Privacy-Preserving Aggregation Scheme for Multidimensional Data in IoT," in IEEE Internet of Things Journal, vol. 9, no. 1, pp. 589-600, 1 Jan.1, 2022, doi: 10.1109/JIOT.2021.3083136.

[4] Yange Chen, Qingqing Zhao, Pu Duan, Benyu Zhang, Zhiyong Hong, Baocang Wang, Verifiable privacy-preserving association rule mining using distributed decryption mechanism on the cloud, Expert Systems with Applications, Volume 201, 2022, 117086, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117086.

[5] Zehtabchi, S., Daneshpour, N. & Safkhani, M. A new method for privacy preserving association rule mining using homomorphic encryption with a secure communication protocol. *Wireless Netw* (2022). https://doi.org/10.1007/s11276-022-03185-5

[6] L. Dias and T. A. Rizzetti, "A Review of Privacy-Preserving Aggregation Schemes for Smart Grid," in IEEE Latin America Transactions, vol. 19, no. 7, pp. 1109-1120, July 2021, doi: 10.1109/TLA.2021.9461839.

[7] Virupaksha, S., Dondeti, V. Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data. *Peer-to-Peer Netw. Appl.* **14**, 1608–1628 (2021). https://doi.org/10.1007/s12083-021-01080-y

[8] H. Pang and B. Wang, "Privacy-Preserving Association Rule Mining Using Homomorphic Encryption in a Multikey Environment," in IEEE Systems Journal, vol. 15, no. 2, pp. 3131-3141, June 2021, doi: 10.1109/JSYST.2020.3001316.

[9] W. A. K. Salman and S. B. Sadkhan, "Privacy Preserving Association Rules based on Compression and Cryptography (PPAR-CC)," 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 2020, pp. 37-42, doi: 10.1109/ICOASE51841.2020.9436603.

[10] Chen-Yi Lin, A reversible privacy-preserving clustering technique based on k-means algorithm, Applied Soft Computing, Volume 87, 2020, 105995, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.105995.

[11] Ezgi Zorarpacı, Selma Ayşe Özel, Privacy preserving classification over differentially private data, Wires Data Mining and Knowledge Discovery, Volume 11, December 2020, e1399, https://doi.org/10.1002/widm.1399

[12] Weibei Fan, Jing He, Mengjiao Guo, Peng Li, Zhijie Han, Ruchuan Wang, Privacy preserving classification on local differential privacy in data centers, Journal of Parallel and Distributed Computing, Volume 135, 2020, Pages 70-82, ISSN 0743-7315, https://doi.org/10.1016/j.jpdc.2019.09.009.

[13] H. Huang, D. Zhang, F. Xiao, K. Wang, J. Gu and R. Wang, "Privacy-Preserving Approach PBCN in Social Network With Differential Privacy," in IEEE Transactions on Network and Service Management, vol. 17, no. 2, pp. 931-945, June 2020, doi: 10.1109/TNSM.2020.2982555.

[14] Xiaodong Zhao, Dechang Pi, Junfu Chen, Novel trajectory privacy-preserving method based on clustering using differential privacy, Expert Systems with Applications, Volume 149, 2020, 113241, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113241.

[15] Lekshmy, P.L., Rahiman, M.A. A sanitization approach for privacy preserving data mining on social distributed environment. *J Ambient Intell Human Comput* **11**, 2761–2777 (2020). https://doi.org/10.1007/s12652-019-01335-w

[16] Clemens Rosner, Melanie Schmidt, Privacy preserving clustering with constraints, Computational Complexity, 16 Feb 2018, arXiv:1802.02497.

[17] Manikandan, V. Porkodi, Amin Salih Mohammed and M. Sivaram, PRIVACY PRESERVING DATA MINING USING THRESHOLD BASED FUZZY CMEANS CLUSTERING, IJSC, Vol 9 ,Iss 1, Paper 6, 10.21917/ijsc.2018.0253

[18] L. Ni, C. Li, X. Wang, H. Jiang and J. Yu, "DP-MCDBSCAN: Differential Privacy Preserving Multi-Core DBSCAN Clustering for Network User Data," in IEEE Access, vol. 6, pp. 21053-21063, 2018, doi: 10.1109/ACCESS.2018.2824798.