Anita Parmar[1]
Rakesh Parmar[2]
Chirag A. Patel[2]

Noise Addition Based Approach for Privacy Preserving Data Stream Classification

Journal of Electrical Systems

ABSTRACT
The research sector is paying close attention to data stream mining, which has many uses in fields including banking, education, networking, telecommunication, weather forecasting, stock market, and more, as a result of the significant increase of massive data streams. As a result, researchers are paying increasing attention to protecting privacy in data stream mining. In this work, we provide a noise addition-based method for data stream privacy-preserving classification that applies classification algorithms to large data streams while maintaining data privacy. This fascinating field of study has recently gained additional insight from the developing big data analytics context.

Key words: classification, data streams, privacy preserving.

1. INTRODUCTION

1.1 Privacy Preserving Data Mining (PPDM)

Numerous parameters can be used by data mining applications to analyze the data. They include clustering (finding and visually documenting groups of previously unknown facts), classification (identifying new patterns), association (patterns where one event is connected to another, like buying a pen and buying paper), and forecasting (finding patterns from which one can make reasonable predictions regarding future activities, like the prediction that people who join an athletic club may take exercise classes).

A significant amount of personal data is routinely gathered and examined due to the growing usage of data mining.Governments and commercial groups use this data extensively in their decision-making processes [1]. However, improper analysis of such data might lead to additional risks to an individual's privacy and autonomy. As a result, the research community focusing on privacy and knowledge discovery has created an intriguing new avenue for data mining research called privacy preserving data mining (PPDM). These algorithms seek to preserve private information while simultaneously extracting pertinent knowledge from massive data collections. In privacy-preserving data mining, there are two primary factors to take into account [2].The first is that personal information, such as names, addresses, and identities, should be kept separate from the original data so that the recipient cannot see the original. Second, information gleaned from data mining algorithms should also be disregarded because it may jeopardize data privacy. Therefore, the primary goal of privacy-preserving data mining is to create algorithms that alter the original data in a way that ensures sensitive information and data stay private even after the data mining process is complete.

Since the year 2000, various PPDM techniques have been created for various uses. PPDM is gaining popularity as a study topic since it allows the sharing of sensitive and confidential data for analysis.

Numerous strategies have been discovered by researchers to prevent unwanted access to sensitive information (or knowledge). The disclosure limitations of sensitive knowledge by data mining algorithms were initially presented by M. Attallah et al. [3], who also suggested heuristic techniques to stop sensitive knowledge from being disclosed.

Some recommendations for defining and assessing privacy preservation were made by the authors in [4]. In [5][6][7], they offered a survey and an outline of data mining techniques that can be used for PPDM of massive amounts of data. This serves as background material for the in-depth explanation of the most popular PPDM techniques that follows.

1.2 Privacy Preserving Data Stream Mining (PPDSM)

During the data mining period, the research sector paid close attention to data stream mining [8]. Additionally, it has a significant impact on a variety of applications, including finance, education, telecommunication, networking, stock market trading, and weather forecasting. As a result, academics are paying more attention to protecting privacy in data stream mining [9].

[1]Research Scholar, Gujarat Technological University,India
[2]Department of Information Technology, L.E.College,Morbi (Gujarat), India.
[1]anitaparmar.it@gmail.com, [2]prof.capatel@gmail.com, [2]rparmar2007@gmail.com

The primary goal of privacy-preserving approaches is to alter data in order to conceal the identify of objects inside it and make it possible to conduct mining operations on the data stream. This alteration can alter the data's initial distribution, which would reduce the data's usefulness for data mining methods. As a result, the stream mining process poses an intriguing issue when privacy and data utility are combined. While privacy-preserving techniques can result in response time delays and make it difficult to detect drift, stream data mining techniques are distinguished by their quick response times and capacity to adjust to changes in data distribution. As a result, the mining model may not be appropriately adjusted.

Depending on the data mining task, the privacy-preserving large data stream mining solutions can be divided into the following categories:

- Data stream publication [10-13]
- Association rule mining [14-15]
- Classification [16-21]
- Clustering [22]

In order to categorize massive data streams while maintaining data privacy, we primarily concentrate on privacy-preserving data stream classification in this study.

2. LITERATURE SURVEY

The literature currently available on data stream classification while maintaining privacy is briefly presented in this section. We emphasize the efforts made to develop classifiers from the stream data that preserve privacy.

The PCDS approach was proposed by Ching-Ming Chao et al. [16] for the classification of data streams while maintaining anonymity. Data stream preprocessing is the first stage of PCDS, followed by data stream mining. They use their suggested DSP method to disturb stream data in the first stage. According to experimental results, the DSP method is more secure than other data perturbation algorithms because its security measurement has lower distance-based record linkage (DBRL) values and greater average squared distance (ASD) values. According to experimental data error measurement results, the DSP algorithm has lower bias in mean (BIM) and bias in standard deviation (BISD) values than other algorithms, indicating higher data utility. Thus, there is less data error in the DSP algorithm.

In [17], an alternative method for classifying data streams while maintaining anonymity was put out. Additionally, their methods consist of two steps: pre-processing date streams and mining data streams. Two algorithms—the rotation perturbation algorithm and the sliding window concept technique—are suggested for data perturbation in the initial data streams pre-processing step. Two metrics are frequently used to evaluate perturbation procedures. The first is the degree of privacy protection, while the second is the degree of data utility protection.

The decrease of classification accuracy is the primary indicator of data utility. To create a perturbed data collection, they used the data perturbation algorithm. The Hoeffding tree algorithm is used to classify the disturbed data stream. Both the original and perturbed data sets' classification models are produced. Classification results are assessed using accuracy metrics. Using the suggested algorithms, the classification result demonstrates data privacy with little information loss on the perturbed data set. Numerical qualities are perturbable via their data perturbation techniques. They presented the P2RoCAl data stream perturbation algorithm in [19], which outperforms comparable techniques in terms of accuracy, efficiency, and attack resilience.

Compared to its competitors, the suggested approach P2RoCAl demonstrated higher classification accuracy. In comparison to rotation perturbation and data condensation, P2RoCAl also exhibits greater resilience against attacks including naïve estimate, I/O attacks, and ICA attacks.

R. Kotecha et al. [18] proposed the Diverse and k-Anonymized HOeffding Tree algorithm, or DAHOT, to preserve output-privacy in data stream classification. This algorithm combines a variation of the k-anonymity and l-diversity principle with the Hoeffding tree algorithm for data stream classification. A DAHOT takes the data stream as input and induces a privacy-preserving decision tree classifier that offers high accuracy given a user-specified anonymity and diversity requirement. Massive data streams can be efficiently classified using DAHOT while maintaining the necessary privacy. Six assessment metrics were used to compare the performance of DAHOT with the Hoeffding tree classifier: information loss, interpretability, training and classification accuracy, and training and classification time. They demonstrated that the classifier generated by DAHOT has minimal information loss, low training time, and great interpretability.

Using Adaptive Random Forest (ARF), Fatlawi et al. [21] suggested a classification model with differential privacy for mining the medical data stream. In their study, they develop and apply an adaptive random forest-based classification model for stream data that incorporates differential privacy. Therefore, there are two primary steps. The first is getting the medical data ready for the mining process. Building an ensemble classifier, which consists of numerous extremely quick decision trees, is the second step. Streaming real batch datasets are used to compare the performance.

They introduced a method for data stream mining and publication that protects privacy in [20]. Random projection, random translation, and additive noise are the two methods of data disturbance. The noise is either generated entirely separately for every record (RPIN) or accumulates over the duration of the stream (RPCN).

Existing techniques for privacy-preserving data mining (PPDM) and privacy-preserving data stream mining (PPDSM) are examined by U. H. W. A. Hewage et al. [23]. They also examined how well the current PPDM/PPDSM techniques handle the trade-off between data privacy and data mining accuracy. Their research indicates that the categorization task is where privacy preservation in stream data is most applicable. They came to the conclusion that PPDSM needs a lot of methods to maximize the trade-off between accuracy and privacy in data stream mining.

According to the research mentioned above, data perturbation and anonymization are the most often employed methods for maintaining privacy in data stream categorization. Table 1 provides a quick overview of related works.

3. PROBLEM DEFINITION

Problem Statement

to convert a specified data stream D into a modified version D' that maintains privacy while sacrificing the least amount of information necessary for the data classification task. Optimizing the accuracy-privacy trade-off in data stream mining is the main concern.

Proposed Framework:

In this case, incremental learning is used. To simulate streaming data, we first divided a huge data stream into smaller pieces. The stream of data is handled in segments. We use a privacy-preserving approach for every chunk. In this case, cumulative noise addition is being used to protect the privacy of the data stream. It simulates streaming settings by gradually adding noise sampled from a Gaussian distribution. The following equation can be used to represent the same thing:

$$Y = X + N(\mu,\sigma^2)\text{where,}$$

Y is Perturbed dataset, X is Original dataset and $N(\mu,\sigma^2)$ is Gaussian noise sampled from a normal distribution with mean μ and standard deviation σ.

A disturbed data stream is produced following the addition of noise. The classifier receives the perturbed data stream, classifies each chunk, and updates the model based on predictions.

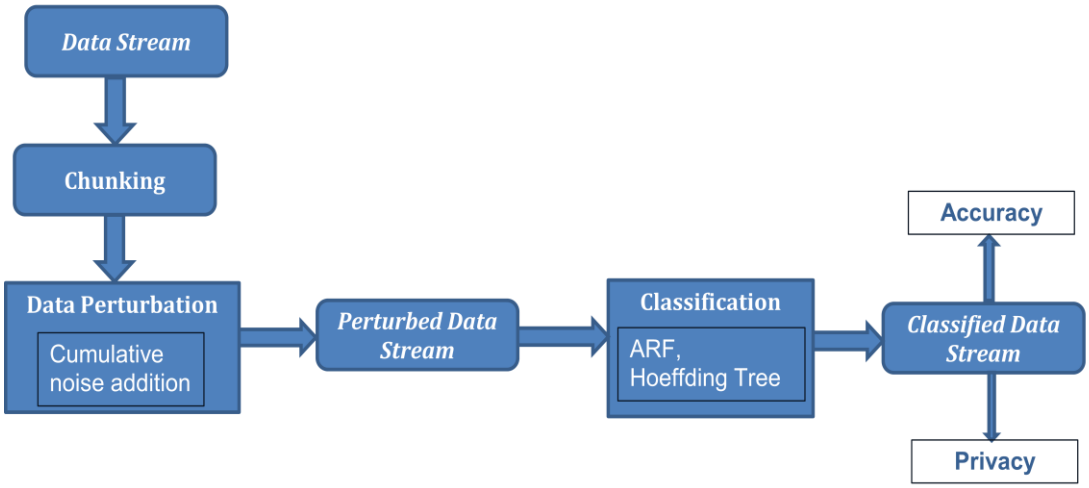

Fig. 1. Proposed framework

We are utilizing the Adaptive Random Forest (ARF) technique for stream data and the Hoeffding tree approach for categorization. A decision tree algorithm called Hoeffding Tree was created for large data streams. It works sequentially, processing one example at a time. After that, it decides which attribute is appropriate for dividing. And nodes are divided appropriately. An ensemble-based learning method called ARF was created for the classification of data streams. It makes use of several decision trees. The ensemble's trees all learn on chunks. A majority vote from every tree is used to make forecasts based on voting procedures.

4. EXPERIMENTS AND RESULTS

Three real-world datasets from the UCI repository were used in our experiment. Each dataset's details are provided in the table.

Dataset	#Attributes	#Instances	#Class
TAXI	7	50000	3
Adult	14	45222	2
Breast Cancer	32	569	2

Two classification algorithms—the hoeffding tree and the adaptive random forest classifier—are used in the experiments. The following are the additional settings for creating noise:

- mean = 0
- variance = 0.0625.
- Cycle size = 2000

RESLUTS:

This section shows the results of the experiments carriedout in given datasets..

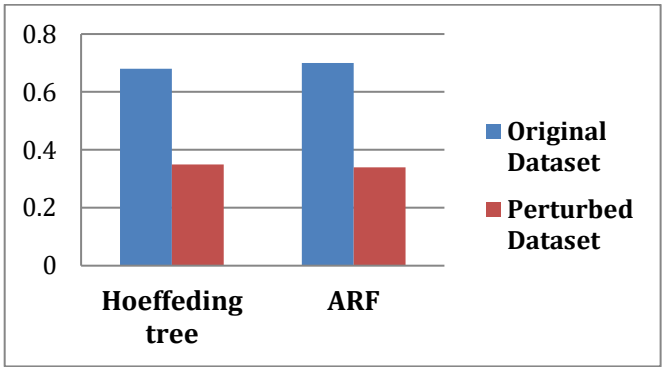

Figure 2. Classification Accuracy(Taxi dataset)

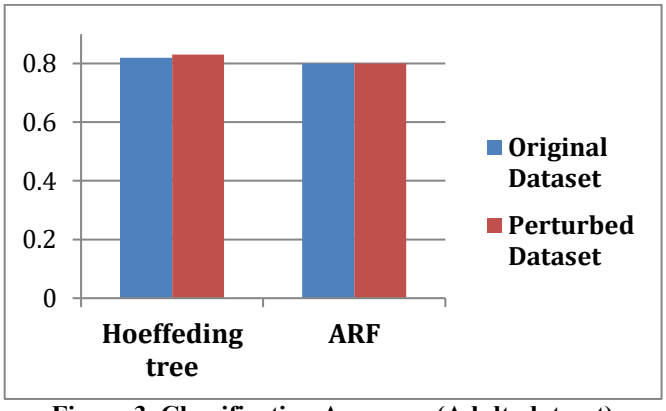

Figure 3. Classification Accuracy (Adult dataset)

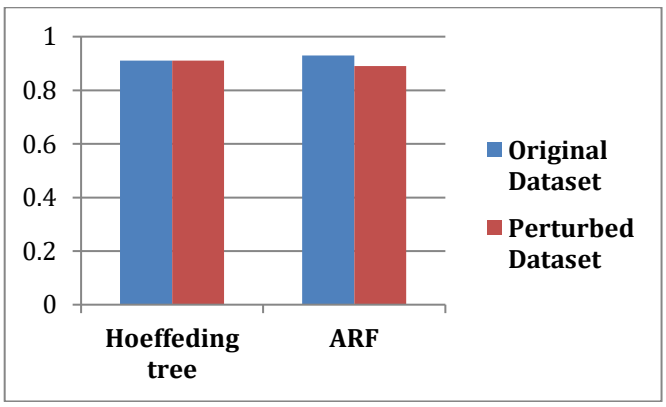

Figure 4. Classification Accuracy (WDBC dataset)

Dataset	Classifier	Original Accuracy	After Perturbation Accuracy
TAXI	HoeffdingTree	0.68	0.344
	ARF	0.71	0.35
Wdbc	Hoeffding Tree	0.91	0.91
	ARF	0.93	0.89
Adult	Hoeffding Tree	0.82	0.80
	ARF	0.83	0.80

We used classification accuracy to assess privacy-preserving classification techniques:

● **Classification Accuracy:** The proportion of cases that the classifier successfully classifies is known as accuracy.

4. CHALLENGES AND DIRECTIONS

Future research efforts can take into consideration a number of research difficulties and directions in the field of privacy-preserving data stream classification. We go over some of these difficulties in the sections that follow.

Problems with Concept Drift.

Concept-drift issues are present in big data streams. This makes meeting privacy-preserving requirements challenging. This is due to the fact that data privacy is maintained based on a predefined set of characteristics of the target data stream.

Privacy Vs Utility.

Big data stream mining methods have conflicting privacy and usefulness features. Determining the appropriate trade-off between these two attributes is therefore a crucial research question. How can privacy be maintained without sacrificingusefulness or quality? Future scientific endeavors should address this question.

Performance.

Processing large data streams while maintaining their privacy always results in performance problems with regard to time, memory, and accuracy. Therefore, creating models that enable us to guarantee the effectiveness of privacy preservation classification techniques across large data streams is a pertinent task for the future.

5. CONCLUSION

We have applied a cumulative noise addition technique to a perturbed data stream in this study. The technique gradually introduces noise into the provided data stream. Following the perturbation, the classifier receives the stream of data and classifies it. We compute classifier accuracy to assess the method's performance.

Even though researchers have worked hard in recent years to classify data streams in a way that preserves privacy, this field is still difficult and offers opportunities to improve on current methods while also creating new, innovative approaches to raise classification accuracy and privacy levels. Big data stream mining techniques have conflicting qualities between accuracy and privacy. Determining the appropriate trade-off between these two attributes is, in fact, a basic research problem.

REFERENCES

1. Jiawei Han and Micheline Kamber ,**Data mining**, second edition, san Francisco, morgan Kaufmann publishers-2006,285-378.
2. Xinjun Qi , Mingkui Zong , **An Overview of Privacy Preserving Data Mining**, Elsevier, Procedia Environmental Science , 12, 2012 .
3. M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, **Disclosure limitation of sensitive rules**, in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999, pp. 45–52.
4. C. Clifton, M. Kantarcioglu, and J. Vaidya, **Defining privacy for data mining**, in National Science Foundation Workshop on Next Generation Data Mining, 2002, pp. 126–133.
5. Ricardo Mendes, Joao P. Vilela, **Privacy-Preserving Data Mining: Methods, Metrics, and Applications**, IEEE Access, 2019
6. Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, Mohammad Abdur Razzaque, **A comprehensive review on privacy preserving data mining**, SpringerPlus (2015) 4:694
7. S.Shimona , **Survey on Privacy Preservation Technique**, (ICICT-2020),CFP20F70-ART,ISBN: 978 - 1 – 7281 – 4685 - 0, IEEE Xplore ,2020.
8. Eiman Alothali, Hany Alashwal, Saad Harous, **Data stream mining techniques: a review**, TELKOMNIKA, Vol.17, No.2, April 2019, pp.728-737
9. Cuzzocrea, Trieste, Italy, **Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions**, 2375-9259/17 , IEEE International Conference on Data Mining Workshops, 2019

10. Gayathri DeviN,Manikandan K,**Improved perturbation techniqueprivacy-preserving rotation-based condensation algorithm for privacy preserving in big data stream using Internet of Things,** Trans Emerging Tel Tech. Wiley, 2020.
11. Jinyan Wang, Chaoji Deng, And Xianxian Li, **Two Privacy-Preserving Approaches for Publishing Transactional Data Streams, special section on recent computational methods in knowledge engineering and intelligence computation**, IEEE, 2018
12. Ganesh Dagadu Puri1 , D. Haritha, **A Novel Method for Privacy Preservation of Health Data Stream Data publishing**, IJATCSE, 2020
13. Bin Zhou1 Yi Han, **Continuous Privacy Preserving Publishing of Data Streams**, ACM, 2009
14. Domadiya N.H., Rao U.P., **A Hybrid Technique for Hiding Sensitive Association Rules and Maintaining Database Quality.** Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2 (2016). Smart Innovation, Systems and Technologies, vol 51. Springer.
15. Jinyan Wang, Chen Liu, Xingcheng Fu, Xudong Luo, Xianxian Li, **A three- phase approach to differentially private crucial patterns mining over data streams, computers & security** 82 (2019) 30–48 , Elsevier, 2019.
16. Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, **Privacy-Preserving Classification of Data Streams**, Journal of Science and Engineering -2009
17. Hitesh Chhinkaniwala, Kiran Patel, Sanjay Garg, **Privacy Preserving Data Stream Classification Using Data Perturbation Techniques**, ICECIT, 2012.
18. R. Kotecha, and S. Garg, **Preserving output-privacy in data stream classification**, Progress in AI 6(2), pp. 87-104, 2017.
19. M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, **Efficient data perturbation for privacy preserving and accurate data stream mining**, Elsevier, Pervasive and Mobile Computing 48 (2018) 1–19
20. Benjamin Denham, Russel Pears, M. Asif Naeem, **Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining**,Elsevier-2020.
21. Fatlawi, Hayder K. and Kiss, Attila. **Differential privacy based classification model for mining medical data stream using adaptive random forest,** Acta Universitatis Sapientiae, Informatica, vol.13, no.1, 2021, pp.1-20.
22. Fang Liu and Tong Li, **A Clustering k-Anonymity Privacy- Preserving Method for Wearable IoT Devices**, WILEY, 2018
23. U. H. W. A. Hewage, R. Sinha,M. Asif Naeem, Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review, Springer-2023