[1] **Anita A. Parmar**
[2]**Chirag A. Patel**
[3]**Rahul R. Keshwala**
[4]**Rakesh A. Parmar**

Privacy Preserving Data Stream Classification: Recent Approaches and Open Challenges

Journal of Electrical Systems

Abstract: - With the relevant growth of big data stream, the research industry has great attention to data stream mining which has a wide range of applications like banking, education, networking, telecommunication, weather forecasting, a stock market, and so on. Because of this, privacy preserving in data stream mining is having more attention from researchers. In this paper, we mainly focus on review of privacy preserving classification methods for data streams, which applies classification algorithms to big data streams while ensuring the privacy of data. Recently, the emerging big data analytics context has conferred a new light to this exciting research area.

Keywords: classification, data streams, privacy preserving, data mining.

I. INTRODUCTION

A. Privacy Preserving Data Mining (PPDM)

All Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), classification (identification of new patterns), clustering (finding and visually documenting groups of previously unknown facts), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).
As the increasing use of data mining, large amount of personal data are regularly collected and being analyzed. This data is an important asset to governments and business organizations both to decision making processes [1]. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly. Thus, an interesting new direction of data mining research, known as privacy preserving data mining (PPDM), has been developed by the research community working on privacy and knowledge discovery. The aim of these algorithms is the extraction of relevant knowledge from large collection of data, while protecting private information simultaneously.
There are two main considerations in privacy preserving data mining [2]. First is personal data like identity, names, address details etc. should be preserved from the original data, so that receiver of the data may not be able to view original data. Second, knowledge gained from data mining algorithm should also be excluded; as such knowledge can also compromise privacy of data. So, the main purpose in Privacy preserving data mining is to develop algorithms which modifies the original data in such a way, so that even after the data mining process sensitive data and knowledge remain private.
Since year 2000 till now, different PPDM methods have been developed for different purposes. As PPDM shares private and sensitive data for analysis purposes, it is becoming a popular research direction.
Researchers have been found many approaches for protecting disclosure of sensitive information (or knowledge) from unauthorized access. M.Attallah et al. [3] were the first to present disclosure limitations of sensitive knowledge by data mining algorithms and proposed heuristic algorithms to prevent disclosure of sensitive knowledge. The authors in [4] presented some suggestions for defining and measuring privacy preservation. In [5][6][7], they provided survey, an overview of data mining methods that are applicable to PPDM of large quantities of information which serves as preliminary background for the subsequent detailed description of the most common PPDM approaches.

B. Privacy Preserving Data Stream Mining (PPDSM)

In the era of data mining, data stream mining got the great attention from research industry [8]. Also it has a great impact on a wide range of applications like networking, telecommunication, education, banking, weather forecasting, a stock market, and so on. Because of this privacy preserving in data stream mining is having more attention from researchers [9].
Privacy preserving techniques aim mostly to make a data modification to hide the identity of the objects in data and enable performing the mining operations on the data stream. This modification may change the original distribution of the data, hence, the effectiveness of data will decrease for data mining techniques. Therefore, the

[1]*Research Scholar, Gujarat Technological University, India*
[2,3,4]*Department of Information Technology, L. E. College, Morbi, Gujarat, India*

combination of privacy and utility of the data for the stream mining process represents an interesting challenge. On the other hand, stream data mining techniques are characterized by fast response and ability to adapt changes in data distribution, while privacy-preserving techniques can cause delays in response time and difficulty in detecting the drift, due to which the mining model may not be adapted properly.

The privacy-preserving big data stream mining technologies can be categorized into following, based on data mining task:

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

In this paper, we mainly focus on privacy preserving classification of data streams, which applies classification algorithms to big data streams while ensuring the privacy of data.

II. LITERATURE SURVEY

This section presents, in brief, the existing literature related to classifying data stream with privacy preservation. We highlight the work carried out to create privacy-preserving classifiers from the stream data.

Ching-Ming Chao et al.[16] proposed the method called PCDS for privacy-preserving data stream classification. PCDS consists of two stages: first step is data streams preprocessing and then second, data stream mining. In the first stage, they perturb stream data using their proposed DSP algorithm. Experimental results show that DSP algorithm has higher security as security measurement of DSP algorithm has higher average squared distance(ASD) values and lower distance-based record linkage(DBRL) values than other data perturbation algorithms. Experimental results of data error measurement showed that bias in mean(BIM) and bias in standard deviation(BISD) values of the DSP algorithm are lower than other algorithms, which indicates higher data utility. So, DSP algorithm has less data error.

Another approach for privacy-preserving classification of data streams was proposed in [17]. Their algorithms also has two steps: date streams pre-processing and data streams mining. In first data streams pre-processing step, two algorithms are proposed for data perturbation that are using sliding window concept algorithm and using rotation perturbation. Many times perturbation techniques are evaluated with two metrics. One is level of privacy preserved and second is level of data utility preserved. Data utility is mostly measured by the loss of accuracy for data classification. They applied data perturbation algorithm to generate perturbed data set. Perturbed data stream is classified using Hoeffding tree algorithm. They classification model of original data set and perturbed data set are generated. Accuracy parameters are used to evaluate classification results. The classification result shows data privacy with minimal information loss on perturbed data set using proposed algorithms. Their data perturbation algorithms can perturb numeric attributes.

In [19], they introduced data stream perturbation algorithm called P2RoCAl which provides higher accuracy, efficiency and attack resilience than similar methods. The proposed method P2RoCAl showed better classification accuracies than its contenders. P2RoCAl also shows higher resilience against the attacks such as naive estimation, I/O attacks, and ICA attacks, compared to rotation perturbation and data condensation.

For preserving output-privacy in data stream classification, R. Kotecha et al [18] proposed an algorithm named DAHOT, Diverse and k-Anonymized HOeffding Tree, that combines variant of k-anonymity and l-diversity principle with Hoeffding tree algorithm for data stream classification and a DAHOT takes as an input the data stream and induces a privacy-preserving decision tree classifier that provides high accuracy given a user-specified anonymity and diversity requirement. DAHOT is efficient in classifying massive data streams while preserving the required privacy. They compared the performance of DAHOT with Hoeffding tree classifier using six evaluation parameters: training and classification accuracy, training and classification time, interpretability and information loss. They showed that classifier induced using DAHOT has high interpretability, requires less training time and causes minimum information loss.

In [20], they presented approach for privacy-preserving data stream mining and publishing. There are two data perturbation methods which are a combination of random projection, random translation, and additive noise that is either generated completely independently for each record (RPIN), or accumulated over the course of the stream (RPCN). Fatlawi et.al [21] proposed a classification model with differential privacy for mining the medical data stream using Adaptive Random Forest (ARF). In their work, a classification model is designed and implemented based on adaptive random forest for stream data, including differential privacy. Thereby, there are two main stages; the first one is to prepare the medical data for the mining task. The second stage is to build an ensemble classifier which includes many very fast decision trees. The performance is compared based on streaming real batch datasets.

From the above literature we can see that for preserving privacy in data stream classification, the mostly used techniques are data perturbation and anonymization. Related works are briefly presented in Table 1.

III. EVALUATION OF PRIVACY PRESERVING DATA STREAM CLASSIFICATION METHODS

To evaluate different privacy preserving classification methods following parameters can be used.

i) ***Classification Accuracy:***

Accuracy is defined as the percentage of instances that are correctly classified by the classifier.

ii) ***Privacy:***

It is a level offered by a privacy preserving technique, which gives the degree of uncertainty. According to which sensitive information has been hidden, it can still be predicted [10].

iii) Utility of Data:

Ensuring the privacy of data streams while mining big data streams requires modification in original data stream. It may deteriorate the quality of such data as original distribution of data may change.

Table 1. Related work on PPDSM

Sr.	Title	Mining task	Techniques Used for privacy preservation	Parameters and Results
1	Privacy-Preserving Classification of Data Streams	Classification	Data perturbation	Security (ASD = 5) data error = 4 - 20%, Error rate (in accuracy)= 12%
2	Privacy preserving data stream classification using data perturbation techniques	Classification	Data perturbation	Information loss = 2-5%, response time = 0.7 seconds
3	Preserving output-privacy in data stream classification	Classification	K-anonymity and l-diversity principles	Accuracy: 80 – 90%, Time: 2 – 10 s, Information loss: 2 -7%
4	Efficient data perturbation for privacy preserving and accurate data stream mining	Classification	Data perturbation	Accuracy = 75% - 95%, time complexity
5	Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining	Classification	Data perturbation	privacy breach (against known input-output attacks), accuracy = 60 - 80%
6	Differential privacy based classification model for mining medical data stream using adaptive random forest	Classification	Data perturbation	Accuracy = 80 to 90%
7	Continuous Privacy Preserving Publishing of Data Streams	Data publishing	K-anonymity	Information loss
8	Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking	Data publishing	Data perturbation	Privacy Discrepancy
9	A Clustering k-Anonymity Privacy-Preserving Method for Wearable iot Devices	Clustering	K-anonymity	Discriminating rate
10	Improved perturbation technique privacy-preserving rotation-based condensation algorithm for privacy preserving in big data stream using Internet of Things	Data publishing	Data perturbation	Privacy Accuracy
11	Novel Method for Privacy Preservation of Health Data Stream Data publishing	Data publishing	L-diversity	Data loss
12	A three- phase approach to differentially private crucial patterns mining over data streams	Pattern mining	Data perturbation	Accuracy (F-score) Data utility (Relative error)

IV. CHALLENGES AND DIRECTIONS

Privacy-preserving data stream classification is a research area which has several research challenges and directions which can be considered by near-future research efforts. In the following, we discuss on some of these challenges.

A. ***Concept-Drift Issues.***

Big data streams have concept-drift problems. This makes difficult to provide privacy preserving requirement. That is because preserving the privacy of data is performed in dependence on a predetermined set of attributes of the target data stream.

B. ***Privacy Vs Utility.***

Privacy and utility are conflict properties for big data stream mining algorithms. So, determining the correct trade-off between these two properties is a fundamental research issue. How to provide privacy while preserving quality and utility? It is a question for future research activities.

C. ***Performance.***

Performance issues in terms of time, memory and accuracy always arise when processing big data streams while preserving their privacy. As a consequence, relevance challenge for the future is to devise models that allow us to ensure performance of privacy preservation classification methods over big data streams.

V. CONCLUSION

The issue of privacy preserving data mining has been widely studied to preserve data privacy during data mining and many techniques have been proposed. However, existing techniques for traditional static data sets are not suitable for data streams. So the privacy preservation in data streams mining is a need for the time.

The research industry has given great attention in solving data streams mining issues like mining of such huge amount of data, suitable algorithm selection, concept drift, skewed data, and performance in terms of time, accuracy and memory while preserving privacy. In this paper, we surveyed privacy preserving classification methods for big data streams.

Despite recent years researchers have put a lot of efforts into privacy-preserving classification of data stream, this field is still certainly challenging and it leaves a room for improving the existing approaches as well as developing new novel approaches to increase classification accuracy and privacy level. Accuracy and privacy are conflict properties for big data stream mining algorithms. Indeed, determining the correct trade-off between these two properties is a fundamental research issue.

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 Devi N, Manikandan K, Improved perturbation technique privacy-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