TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

A Survey on Privacy Preserving Clustering Analysis in Big Data Environment

[1]Dr. Manjunath T N, [2]Amogh Pramod Kulkarni, [3]Dr. Ravindra S Hegad, [4]Prabhuram
[1] Dean-ER & Prof. Dept. of ISE, BMS Institute of Technology and Management, Bengaluru, Research supervisor VTU, Belagavi.
[2] Asst. Prof. Dept of CSE, Presidency University, Bengaluru Research scholar VTU, Belagavi.
[3]Dept. of CSE, Central University, Kalaburgi
[4]Principal Test Lead Consultant, Numerify

Article Info
Volume 83
Page Number: 217 - 223
Publication Issue:
May - June 2020

Article History
Article Received*: 11August 2019*
Revised*: 18November 2019*
Accepted*: 23January 2020*
Publication*:07May2020*

Abstract:
Big data computing has gained wide acceptance for its capability to mine knowledge from a large volume of data. It has been used in many knowledge mining requirements in various domains like medicine, finance, social network analysis etc. Clustering is one of the most common important methods for knowledge extraction from large volumes of data. Mining on data in domains like medicine, finance and social network does compromise the privacy of the individual and often leaks sensitive information. The leak of sensitive data can be direct or through inference. Many methods have been proposed in literature for privacy preservation during data mining. This work studies those methods and identifies the weakness in those solutions when applied for big data analytics.

I. INTRODUCTION

Rapid adoption of digitization has increased the volume of data accumulated in enterprises. These data has wealth of knowledge hidden in it and it can be analyzed to extract patterns related to customer behavior, customer interests etc. The extracted knowledge can be used for target specific marketing, advertisements, product design etc. Prediction, forecasting and recommendation are typical tools in data analytics applications. Some of the examples data analytics applications listed below

Social Network - Friends suggestions and new feed suggestion in Social Networking sites like Facebook.
E-commerce - Product are recommended in ecommerce sites like Amazon, Flipkart.
Business - Data analytics is used for many strategic decisions in areas of customer retention, customer satisfaction in companies like CoCo-Cola.
Travel - Based on history of user movement hotels and travels plans are recommended bysites like Makemytrip.

These data collected by enterprises has many private and sensitive data which could comprise the privacy and security of the people involved. Leakage of gender, medications, user movement information, demographic information, caste, religion associations etc compromises user privacy and attackers can use this information for malicious purposes. The most common privacy threats in data analytics are listed below
Monitoring of transactions.
Private information leakage.

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

Hidden pattern inference.
Abuse.
Privacy preserving data mining has gained increasing attention by researchers. Many privacy preserving techniques have been proposed in literatures to protect against privacy threats. The existing techniques for privacy preservation can be grouped in following categories
Anonymization.
Randomization.
Cryptographic techniques.
Diversification.
Aggregation.
The privacy preserving data mining methods proposed in this category lacks certain properties related to scale of operations when applied to big data environment. There has been many surveys on privacy preserving data mining techniques but not much work has been done for the privacy preserving clustering in big data environments for case of both stored and streaming data. With increased application on big data analytics in many enterprise, this study becomes important. This motivates us to study the existing solutions on privacy preserving data mining related to clustering and suitability of big data environment. This study identifies the open issues and documents it for further research on solution design.

The number of attacks has been rampant in last few years. Some of the major data breaches occurred and leakage data statistics in recent years ([20] Identity Theft Resource Center) is given below:

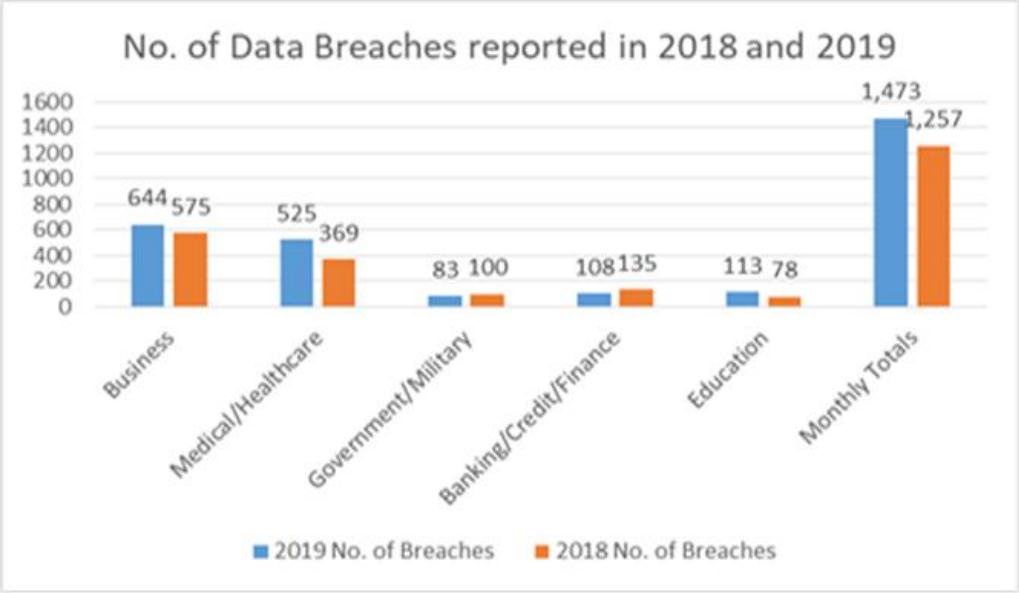

Figure 1: Comparision of data breaches reported in 2018 and 2019 across different sectors

The figure 1 clearly shows there is an approximately 17% rise in reported data breaches in 2019 compared to 2018. While the breaches in government/military and Banking/credit/finance data has reduced, the other sectors have seen rise in data breaches among them medical/healthcare data breaches have the maximum increase.
Moving forward in the development of smart cities, IoT technology plays a major role. There is a need for considering privacy protection as a major criteria in development of IoT solutions. The Figure 2 shows Eclipse IoT developers survey [21]. Security, Connectivity, data collection and analysis, performance, privacy, integration with hardware, standards, and return on investment are the concerns for IoT developers. This survey shows that concern towards privacy has increased to 18% in 2019 as compared to 11% in 2018. It is still unfortunate to

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

note that concern towards privacy is at fourth position among IoT developers. There is need among the developers to prioritize privacy protection as a major concern. The developers must build solutions that provides and protects privacy of users.

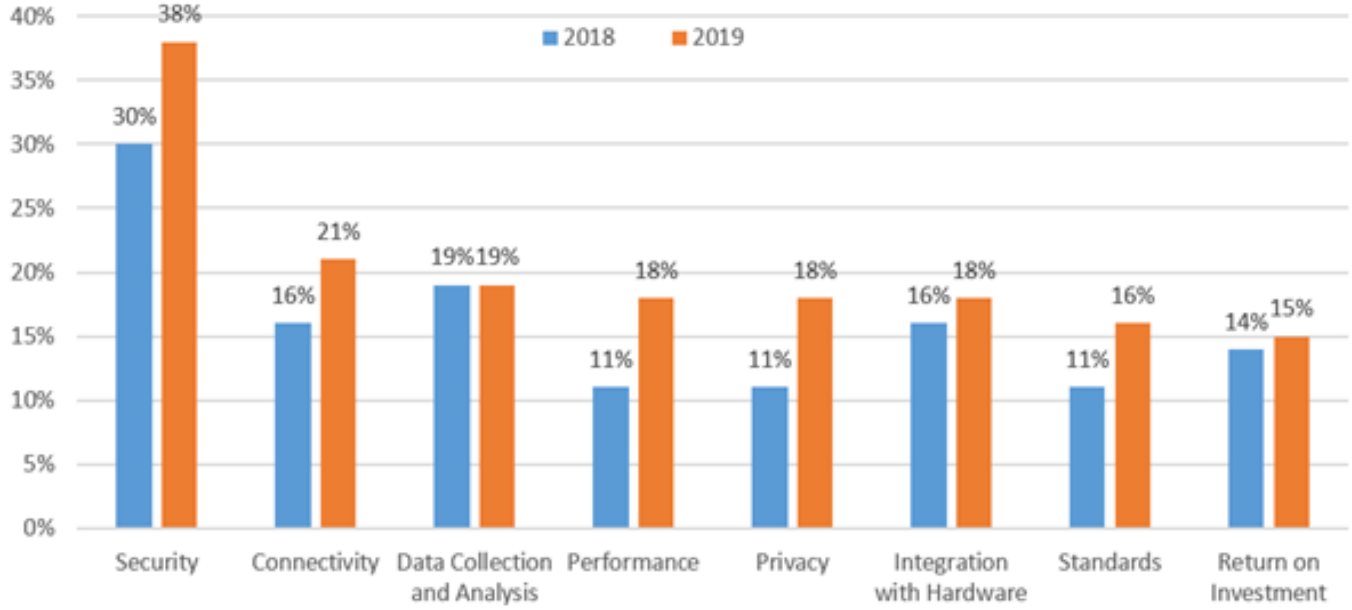

Figure 2: Survey of top concerns for developing IoT solutions

Policymakers have a significant part to play in this. Organizations that have been infringed need incentives and/or safeguards that allow them to share more information about infringements without fear of complicating legal and regulatory proceedings. Likewise, policymakers and regulators should continue to develop framework/systems at all levels to promote best practices in data collection, usage, storage, destruction, and overall cyber security.

The customer usually asks for convenience over a more secure process. This creates an environment where bad actors have less obstacles to achieving their goals. Consumers must choose if the added ease of quicker access outweighs the increased security offered by additional measures. The consumers must understand that for the organizations we work with, to demand better security and privacy is not an unusual proposition.

II. Survey

A. Privacy Footing and Techniques

"The definition of privacy is an individual's right to control access to her/his personal data, such as identifying information, bio-specimen and so on" [1]. The main goal is to provide readers with an opportunity to consider what has been achieved (mathematical concepts, algorithms, frameworks, and proposed solutions), what needs to be tackled, and the challenges facing privacy protection. For the sake of handiness, we summarize the progress in the domain of privacy preserving clustering in Table I.

Table 1: Summary of research literature

Ref. No.	Advantages	Problems
[2]	Customizable privacy protection, considered both historical and future data in form of data streams.	Works only on data types with discretized values and fails for continuous values.
[3]	Due to use of Homomorphic encryption, the operations needed for clustering like distance	Time complexity is high for applicability in big data environment.

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

	computation etc is not affected.	
[4]	Locality sensitive hashing on encrypted data for grouping similar objects.	Clustering can be done only on a particular data attribute.
[5]	Data masking and concealment is done in proportional to the sensitivity level.	Order preserving property is violated so cannot be used for clustering.
[6]	Differential privacy preservation.	Time complexity is high for big data.
[7]	Model protection using a multivariate polynomial approach.	Not applicable for clustering models
[8]	Long Short Term Memory (LSTM) Encoder Decoder ensures strong privacy.	Computational complexity grows exponentially for big data
[9]	Encoder/decoder system for privacy preservation in videos.	Concept must be extended for textual data.
[10]	The idea of human interaction in process of anonymization is found to yield better results.	The approach cannot be applied in same manner in big data environment.
[11]	User can configure the best assembly for anonymization with a goal of reducing information loss.	The approach cannot be applied in same manner in big data environment.
[12]	Sharp tradeoffs between privacy protection and estimation rates for statistical models.	Not applicable for clustering models.
[13]	Hybrid approach that combines Top–Down Specialization (TDS) and Bottom–Up Generalization (BUG).	Locality sensitivity is lost.
[14]	Top-down specialization (TDS) approach for big data based anonymization.	Lacks control on level of anonymization.
[15]	Efficient topic-based clustering of encrypted unstructured big data.	Only works for certain encryption methods.
[16]	Cryptography free and based on multiparty additive scheme.	Works only for multi-party horizontal portioned data.
[17]	Able to cluster the user's data into correct clusters without knowing any useful information about the model and user data.	The major limitation is this solution is that its model cannot be retrained with new data due to which accuracy over a period drops.
[18]	Nearest similarity based clustering (NSB) with Bottom-up generalization based on sensitivity measure.	The sensitivity measure is specific to dataset.

B. Government of India Perspective towards Data Privacy

The right to privacy has been recognized as a fundamental right emerging primarily from Article 21 of the Constitution of India. To make this fundamental right effective, it is the State's responsibility to create a data protection system that serves the common good. The data protection system should defend citizens from dangers to personal privacy that originates from state and non-state actors. It is this understanding of the state‘s duty that government of India in December 2017 constituted an expert committee. The objective of the committee was to study and find key data protection issues and

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

mention techniques to address them. A white paper was released by expert committee for public consultation. The Government of India has come up with Personal Data Protection Bill in the year 2019. This PDP bill after getting passed in Indian parliament will be enacted as PDP Act. This PDP Act would regulate how data of the country's citizen is captured, stored, analyzed and transferred.

Prominent points of the PDP Bill includes the formation of a data protection authority, necessitates technology companies to obtain explicit permission for use of personal data and allowing citizens more rights over their personal data [19]. It enables the central government to immune government agencies from the bill's requirements "in the interest of sovereignty and integrity of India". Under the bill, intermediaries on social media would be needed to provide users with an opportunity to verify their identity. Additionally, it provides both the right to data expurgation and the right to be forgotten, regulates research on data, and deeply regulates biometrics. The bill also describes penalties and remedies for violation or noncompliance of policy guidelines.

The bill clearly defines personal data and sensitive personal data with varied level of protection guidelines for them. The PDP bill clearly distinguishes grounds for processing of personal data, from grounds for processing sensitive personal data. The difference in policy guidelines between personal data and sensitive personal data extends for data storage limitation, restrictions on cross-border transfer of data, condition on cross-border transfer of data. The other categories of data defined under the PDP bill are financial data, biometric data, genetic data, and health data. The PDP bill mandates privacy of personal data is protected throughout processing from the point of collection of data to the point of deletion. Also it provisions legitimate business goals including any creativity, are met without violating the values of privacy.

As more and more countries are realizing the importance of data privacy the companies are becoming transparent and are changing their data privacy policy to comply with prevailing law of the land in specific and also in general globally. Twitter has a new global privacy policy that comes into effect from 1st January 2020. In this policy twitter has made users aware of their data being collected and the use of collected data. Also the user can have control over both. This has empowered the users to make correct choice about the data that the end users share on the platform.

III. Issues

The open issues in the existing solutions, according to the referred literature for privacy preserving clustering in big data environment are listed below.

Lack of models for automatic security level calculation.

Deep learning models for privacy preservation have not been explored for clustering.

Interactive privacy preservation with integration of human knowledge is not explored for big data environment

IV. Discussion on Issues

Issue 1: In approaches like [2], the sensitivity of the data is decided by the user. But in big data environment with multi attribute dataset especially for data generated by IOT devices, there should be automatic methods to calculate the sensitivity of the data based on its attribute distribution and statistical properties. This area is not much explored.

Issue 2: Deep learning-based encoder/ decoders have become a recent trend for data obfuscation or transformation. These methods perform best against attacks but the problem in these approaches is time complexity. Low complexity encoders with order preserving property will help for efficient clustering in big data environment.

Issue 3: Interactive privacy preservation is proved to achieve better anonymization in literature.

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

But the current approaches are not scalable for big data environments. Integration of rule based or expert system-based reasoning for interactive privacy preservation will be efficient for anonymization of large volume of datasets.

V. Conclusion

The paper summarizes the current works in privacy preserved clustering in big data environment. The existing solutions have been detailed and the problems in each solution are documented. The open areas for further search are listed with discussion on the issue. Also the paper summarizes the Government of India, personal data protection bill guidelines. The paper emphasizes that there is a need for all the stake holders' such as developers, Policymakers and consumers commitment towards privacy protection. Further work will be on design on efficient solutions to address the identified open issues.

References

[1] Shuang Wang, Luca Bonomi, Wenrui Dai, Feng Chen, Cynthia Cheung, Cinnamon S. Bloss, Samuel Cheng, Xiaoqian Jiang, "Big Data Privacy in Biomedical Research," IEEE TRANSACTIONS ON BIG DATA no. 99, pp. 1-1, 2016.

[2] Yang D, Bingqing Q, Cudre-Mauroux P. "Privacy-preserving social media data publishing for personalized ranking based recommendation," IEEE Trans Knowl Data Eng. 2018. ISSN (Print):1041-4347, ISSN (Electronic):1558-2191.

[3] Liu Y et al. "A practical privacy-preserving data aggregation (3PDA) scheme for smart grid," IEEE Trans Ind Inf. 2018

[4] Jiang R, Lu R, Choo KK. "Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data," Future Gen Comput Syst. 2018;78:392–401.

[5] Al-Zobbi M, Shahrestani S, Ruan C. "Improving MapReduce privacy by implementing multi-dimensional sensitivity based anonymization," Journal of Big Data. 4. 10.1186/s40537-017-0104-5.

[6] Shang T et al. "A DP Canopy K-means algorithm for privacy preservation of Hadoop platform," International symposium on cyberspace safety and security. Cham: Springer; 2017.

[7] Jia Q et al. "Preserving model privacy for machine learning in distributed systems," IEEE Trans Parallel Distrib Syst. 2018; 29(8).

[8] Psychoula I et al. "A deep learning approach for privacy preservation in assisted living," arXiv preprint arXiv :1802.09359. 2018.

[9] Feng Dai, Dongming Zhang, and Jintao Li. "Encoder/decoder for privacy protection video with privacy region detection and scrambling," International conference on multimedia modeling. Springer , pages 525–527, 2013.

[10] Bernd Malle, Peter Kieseberg, Andreas Holzinger, "Interactive Anonymization for Privacy aware Machine Learning," IAL@PKDD/ECML 2017

[11] Carlos Moque, Alexandra Pomares, and Rafael Gonzalez. "AnonymousData.co: A Proposal for Interactive Anonymization of Electronic Medical Records," Procedia Technology, 5:743–752, 2012.

[12] Martin J Wainwright, Michael I Jordan, and John C Duchi. "Privacy aware learning," Advances in Neural Information Processing Systems, pages 1430–1438, 2012

[13] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud," Journal of Computer and System Sciences, 80(5). pp. 1008-1020, 2014.

[14] Xuyun Zhang ; Laurence T. Yang ; Chang Liu ; Jinjun Chen, "A scalable two phase top-down specialization approach for data anonymization using map reduce on cloud". IEEE Transactions on Parallel and Distributed Systems 25(2), pp. 363-373, 2014

[15] Zobaed, Sm & Ahmad, Sahan & Gottumukkala, Raju & Salehi, Mohsen. (2019). "ClustCrypt: Privacy-Preserving Clustering of Unstructured Big Data in the Cloud," 10.1109/HPCC/SmartCity/DSS.2019.00093.

May – June 2020
ISSN: 0193-4120 Page No. 217 - 223

[16] Z. Gheid and Y. Challal, "Efficient and Privacy-Preserving k-Means Clustering for Big Data Mining," 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, 2016, pp. 791-798.doi: 10.1109/TrustCom.2016.0140.

[17] Hui Yin,, Jixin Zhang,, Yinqiao Xiong "PPK-Means: Achieving Privacy-Preserving Clustering Over Encrypted Multi-Dimensional Cloud Data," MDPI , Nov 2018

[18] P. Srinivasa Rao & S. Satyanarayana "Privacy preserving data publishing based on sensitivity in context of Big Data using Hive," Journal of Big Data volume 5, 2018.

[19] Ministry of Electronics and Information, Government of India, Available at: https://meity.gov.in/writereaddata/files/Personal_Data_ Protection_Bill,2018.pdf (Accessed February 1, 2020).

[20] Identity Theft Resource Centre, Available at: https://www.idtheftcenter.org/wp-content/uploads/2020/01/01.28.2020_ITRC_2019-End-of-Year-Data-Breach-Report_FINAL_Highres-Appendix.pdf (Accessed January 30, 2020).

[21] Available at https://iot.eclipse.org/resources/iot-developer-survey/iot-developer-survey-2019.pdf