International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: 25 May 2023 Revised:12 June 2023 Accepted: 30 July 2023

Hybrid Approach for Data De-Identification during the Storage Phase in Big Data

Tanveer Ahmad Dar, Dr. Surendra Yadav
Department of Computer Science and Engineering
Vivekananda Global University Jaipur Rajasthan
tanveerabdullah786@gmail.com and surendra.yadav@vgu.ac.in

Abstract: With the exponential growth of big data, protecting sensitive information during storage has become a significant challenge. Data de-identification techniques, such as k-anonymity, l-diversity, and t-closeness, ensure privacy by anonymizing identifiable attributes. However, these methods often result in a trade-off between privacy and data utility. This paper proposes a hybrid approach combining Genetic Algorithms (GA) and Simulated Annealing (SA) to optimize data de-identification during the storage phase. The proposed framework balances privacy preservation with minimal information loss, making it suitable for secure storage in large-scale datasets. Experimental results demonstrate the hybrid approach's effectiveness in enhancing privacy while maintaining high data utility for subsequent analytics.

***Key words*:** Big data, Privacy, Security and hybrid approach

1. Introduction

The digital era has led to an unprecedented surge in the volume of data collected, stored, and analyzed across various domains such as healthcare, finance, and social media. While big data offers immense opportunities for insights, it also raises concerns about privacy breaches, particularly during the storage phase. Sensitive information can be misused if not appropriately protected, necessitating robust de-identification techniques.De-identification involves modifying data to ensure individual privacy while retaining utility for analysis. Techniques such as k-anonymity generalize or suppress data to prevent re-identification, but higher levels of de-identification can lead to significant information loss. This challenge underscores the need for advanced optimization techniques to balance privacy and utility effectively.

This study proposes a hybrid optimization framework combining Genetic Algorithms (GA) and Simulated Annealing (SA) for data de-identification during the storage phase in big data. The hybrid approach optimizes parameters for de-identification techniques, reducing information loss and enhancing the usability of anonymized data.

2. Challenges in Data De-Identification

2.1 Privacy vs. Utility Trade-off

- Higher privacy levels (e.g., larger k values in k-anonymity) often lead to increased information distortion.
- Excessive generalization or suppression may render data unsuitable for meaningful analytics.

2.2 Scalability in Big Data

- Traditional de-identification methods struggle with the volume, velocity, and variety of big data.
- Real-time optimization for privacy preservation during storage is critical but computationally expensive.

2.3 Security Risks in Storage

- Inadequately anonymized data during storage phases remains vulnerable to attacks, re-identification, or unauthorized access.
- The specification of Privacy insurance policies managing to get the right of entry
- The technology of productive enforcement video display units for these policies, and
- The integration of the generated video display units into the goal analytics platforms.

Enforcement strategies proposed for usual DBMSs show up insufficient for the massive records context due to the strict execution requirements wanted to deal with giant statistics volumes, the heterogeneity of the data, and the velocity at which records should be analyzed

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: 25 May 2023 Revised:12 June 2023 Accepted: 30 July 2023

3. Challenges: Security and Privacy in Big Data

Security center of attention on shield the enterprise. Data privacy focuses on user's information. There are broadly speaking three targets of security are secrecy, reliability and accessibility. As indicated by way of the article with the aid of Cloud Security Alliance (CSA) [8], there are especially many difficulties in the discipline of Big Data security and protection as referenced beneath:

1) Secured calculations in dispersed programming systems
2) Security great practices for non-relational facts stores
3) End-point enter approval and sifting
4) Real-time protection observing
5) Privacy-safeguarding records mining and examination
6) Cryptographically upheld facts pushed security
7) Granular get right of entry to control
8) Secure records stockpiling and exchanges logs
9) Granular
10) Data proven

3.1 Privacy Protection In Big Data

Privacy is principal challenge in big data so we want environment friendly privacy renovation methods. Privacy without delay related to customers. Privacy normally focuses on user's data as a substitute than complete series of data. The privacy preservation methods can be used defend the individuals sensitive information. Privacy is vital in three tiers i.e. information generation, data storage, information processing. In this paper is focusing on a hybrid approach to maintain the privacy and security to the data during the storage phase

3.2 Big Data Privacy In Data Storage Phase

Big Data stores huge quantity of data. There are precise strategy to keep privacy in storage. Security consists of especially three dimensions i. e. Confidentiality, integrity, and availability [9]. Cryptographic encryption mechanisms are public key encryption, identification particularly based encryption, attribute primarily based encryption etc. In public key encryption approves an data sender to scramble the facts below the general public key of recipient, the recipient decrypts the statistics underneath personal key recipient. So, there can be leakage of data. This cryptographic mechanism does now no longer fulfill every one of the stipulations of consumers in the state of affairs of correct sized facts stockpiling.In an ordinary encryption mechanisms can't approval the anonymity of cipher textual content of receiver /sender. So, absolutely everyone can without difficulty acquiring a cipher textual content (e.g. cloud server), if anyone is aware of the public key of the discern message, that is scrambled below the proprietor of the discern content material. So, outsider can besides a lot of a stretch receives the undeniable text [10].

4. Proposed Approach to Provide Privacy and Security to the Data

In this paper a hybrid approach is proposed to provide the privacy and security to the data during the storage phase of the big data. The diagram and working of the proposed approach is as follows:

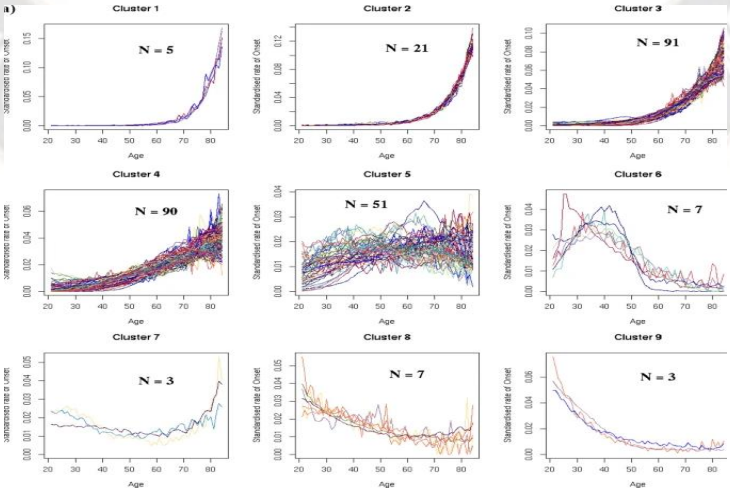

Fig 1: Proposed Approach

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: 25 May 2023 Revised:12 June 2023 Accepted: 30 July 2023

Step 1: Choose one of the characteristic/ attribute or identifier available the dataset like year of birth, age, Licence number, voter id number etc.,
Step 2: Sort them in any order (increasing or decreasing).
Step 3: Construct a tree by choosing the smallest two among the sorted list. Step 4: Repeat step 3, until root node (or single node) is reached.
Step 5: Mark 0 (zero) on the left side and 1 (one) on the right side of the tree.
Step 6: Obtain the unique code of each attribute, by following from root to the attribute.

Illustration of the proposed approach through example

S. No	Pin Code	Age	District	Disease
1	192101	46	Anantnag	Blood Pressure
2	192231	44	Kulgam	Heart Problem
3	192301	36	Pulwama	Tumor
4	192231	35	Kulgam	Tumor
5	192101	45	Anantnag	Heart Problem
6	190002	56	Srinagar	Cancer
7	192301	52	Pulwama	Arithitus

Table 1: Dataset used to illustrate the proposed approach

In this example the attribute age is chosen to de-identification for data, are sorted in increasing order (35, 36, 44, 45, 46, 52, 56), construct a tree and obtain the unique code for each attribute as shown in below table

S. No	Pin Code	District	Codes
1	192101	Anantnag	110
2	192231	Kulgam	100
3	192301	Pulwama	001
4	192231	Kulgam	000
5	192101	Anantnag	101
6	190002	Srinagar	011
7	192301	Pulwama	111

Table 2 : Show the data in de- identification code of the dataset.

Column code of table 2 is obtain by de-identification code of two columns table 1 (age and disease). The proposed study seems to be the best approach for data storage in big data and may provide the greater performance as far as the privacy and security is concerned in big data

4.1 Privacy- preserving big data publishing

The publication and dissemination of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time [11]. Privacy-preserving models broadly fall into two different settings, which are referred to as input and output privacy. In input privacy, the primary concern is publishing anonymized data with models such as k-anonymity and l-diversity. In output privacy, generally interest is in problems such as association rule hiding and query auditing where the output of different data mining algorithms is perturbed or audited in order to preserve privacy. Much of the work in privacy has been focused on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently [12]. Despite the fact that k-anonymity can prevent identity attacks, it fails to protect from attribute disclosure attacks because of the lack of

diversity in the sensitive attribute within the equivalence class. The l-diversity model mandates that each equivalence class must have at least l well-represented sensitive values. It is common for large data sets to be processed with distributed platforms such as the Map Reduce framework [13, 14] in order to distribute a costly process among multiple nodes and accomplish considerable performance improvement. Therefore, in order to resolve the inefficiency, improvements of privacy models are introduced.

5. Conclusion

Big data is analyzed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analyzed. Privacy and security is the major concern of big data and requires higher computational strength, Privacy is concerned with the privilege to have control over the data collected and used by the person. Security is concerned to the execution of protecting data and its benefits through the use of technology by the processes and training from unauthorized access, leak. In this paper, we have investigated the privacy and security challenges in big data by proposing a hybrid approach to De- identify the data in big data during its storage.

In future, the same concept should be elaborated to decrypt the data at receiver's side after the data transmission over the network.

References

[1] Samarati, P., & Sweeney, L. "Generalizing data to provide anonymity when disclosing information." Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998.

[2] Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[3] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. "Optimization by Simulated Annealing." Science, 220(4598):671-680, 1983.

[4] Machanavajjhala, A., et al. "l-Diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data, 2007.

[5] Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. Elsevier, 2011.

[6] Machanavajjhala, A., Kifer, D., Gehrke, J.andVenkitasubramaniam, M., 2007. l-diversity:Privacy beyond kanonymity. ACM Transactions onKnowledge Discovery from Data (TKDD), 1(1), p. 3.

[7] D N Goswami, Anshu Chaturvedi and Mohammad Altaf Dar, "A Generalized Software Reliability Growth Model with different severity of faults" International Journal of Applied Studies, Vol. 3 Issue 11, 2014.

[8] D N Goswami, Anshu Chaturvedi and Mohammad Altaf Dar, "Software Reliability Growth Model with varying-Time fault removal efficiency as well as with fault Introduction" International Journal of Science and Research, Vol. 4 Issue 2, 2015.

[9] Mohammad Altaf Dar, D N Goswami and Anshu Chaturvedi, "Generalized Framework with Different Severity of Faults for Modelling Software Reliability Growth during Testing", International Journal of Advanced Research in Computer Science & Technology, Vol. 3, Issue 1, 2015.

[10] Mohammad Altaf Dar, D N Goswami and Anshu Chaturvedi, "Testing effort dependent Software Reliability Growth Model with dynamic faults for debugging process", International Journal of Computer Applications, Vol. 113, No. 11, 2015.

[11] Mohammad Altaf Dar, Showkat Ahmad Teeli and Fayaz Ahmad Bhat, Framework For Modelling Software Reliability Growth For Error detection With Dynamic Faults", International Journal of Advanced Scientific Research and Management, Volume 3 Issue 9, Sept 2018

[12] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[13] A Cloud Securty Alliance Collaborative research, "Expanded Top Ten Big Data Security andPrivacy challenges" , April 2013.

[14] Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage Kaitai Liang, Willy Susilo, Senior Member, IEEE, and Joseph K. Liu 2015.

[15] Privacy Preservation in the Age of BigData :A SurveyJohn S. Davis II, Osonde A. Osoba