ISBN: 978-1-948012-16-4

Contemporary Issues in Computing (CIC)

DOI: <http://doi.org/10.26480/cic.01.2020.126.129>

A CRYPTOGRAPHIC APPROACH FOR PRIVACY PRESERVING DATA MINING FOR BUSINESS ENVIRONMENT

Bhanu Singh Chouhan, Divyanshu Sharma, Vasudha Yadav, Hitesh Kag

Computer Science Department, MediCaps University Indore, Madhya Pradesh, India

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 26 October 2020
Accepted 27 November 2020
Available online 03 December 2020

ABSTRACT

In a number of business scenarios the data collaboration and analysis is essential for making critical decisions. But all the business owners are worried about end client privacy and data security. Therefore the proposed work addresses this issue and proposed to design a privacy-preserving data model. That model enables a business owner to combine their data to other business owner's data and mine the data patterns on entire data. Thus first a cryptographic technique based on the AES algorithm is implemented and then the ciphered data is used on the server for generating the decision rules. Finally, the rules are distributed to all the target parties. And the parties can only recover the data which are contributed by self. The implementation of the proposed secure data mining model is carried out using the JAVA and WEKA technology. The experimental results demonstrate the proposed technique outperforms in terms of Accuracy, memory and time complexities.

KEYWORDS

privacy preserving data mining, decision tree, decision rules, IF-THEN-ELSE, privacy and security.

1. INTRODUCTION

Data mining is an essential technique of extracting the application based target patterns from raw and huge data. These techniques are usages computational algorithms for performing the required task (Deepashri). According to literature, data mining techniques can be supervised or unsupervised in nature. The supervised learning techniques first accept initial predefined samples or examples to learn the patterns and then the unidentified patterns are recognized on the basis of previous learning (Bhamare et al., 2019). On the other hand, the unsupervised learning algorithms are directly applicable to the data and automatically create groups of data according to the data internal similarity or differences. However, supervised learning methods are much accurate as well as efficient with respect to unsupervised learning algorithms (Mahdaviinejad et al., 2018).

In this context, data mining techniques are much helpful in various real-world applications such as prediction, classification, association, and others (Kavakiotis et al., 2017). But sometimes for mining and extracting patterns from data can be sensitive because of the security and privacy point of view. In order to understand these issues let us assume that, there are a number of business owners who want to conduct an analysis on the consumer behavior for a target industry. In this scenario, a number of different business owners need to combine their own part of data in a common place for utilizing the data mining services and extraction of meaningful patterns. But the end data owners' records available in the outsource data can create privacy and security issues. Therefore in this presented work, a client-controlled privacy-preserving data mining model is proposed for implementation and design.

This segment makes available a fundamental indication of the proposed privacy-preserving data mining technique. In the next section, the related recent development is reported in privacy-preserving data mining. Further, the proposed data mining model is demonstrated and then the presentation of the realized system is provided. Finally based on the experiments and observed facts the conclusion and future work is proposed.

2. LITERATURE SURVEY

This section provides the recent development and contributions of different researches for finding the best and efficient approach for designing the privacy-preserving data mining system.

The compilation and assessment of information are expanding because of the event of processing gadgets. The examination of data is encouraging organizations and contributing helpfully to society in numerous fields. In any case, this stockpiling and stream of touchy information present genuine concerns. Techniques that permit information extraction, while saving protection, are known as security safeguarding information mining (PPDM). R. Mendes et al reviews the important PPDM strategies and the measurements used to assess such methods and uses of PPDM. Besides, the present difficulties and open issues are talked about.

Protection safeguarding is basic where information mining is changed into a helpful errand. V. Manikandan et al propose a security saving edge grouping that utilizes a code-based system with the edge for sharing of mystery information in the protection safeguarding instrument. The procedure empowers the data to be apportioned into various offers and dealt with freely. This technique takes less emphasis in correlation with existing strategies that not require any trust among the customers or

Quick Response Code	This paper was presented at	Access this article online	
	International Conference on Contemporary Issues in Computing (ICCIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/cic.01.2020.126.129

servers. Furthermore, it gives results on the security and proficiency of the technique.

Information irritation is an appreciated information-digging practice for protection safeguarding. A significant issue in information bother is the means by which to adjust the two clashing elements – security and information utility. S. Upadhyay et al proposes a Geometric Data Perturbation (GDP) utilizing information parceling in three-dimensional turns. The properties are separated into gatherings of three and each gathering of qualities is turned about various sets of tomahawks. The turning round edge is picked to such an extent that the fluctuation bolstered security metric is raised which makes the first information recreation muddled. The same number of calculations like arrangement and bunching are invariant to geometric irritation, the information utility is protected right now. The assessment shows that the technique gives great protection safeguarding results and information utility.

The PPDM is assuming a significant job go about as rising innovation to perform different information mining procedures on private information and to give information in a safe manner to ensure touchy information. Numerous kinds of strategy for example randomization made sure about aggregate calculations, and k-namelessness has been recommended to execute PPDM. Rajesh N. et al [4], looks into the PPDM strategy with fluffy rationale, neural system, made sure about entirety, and different encryption calculations. This will permit grabbing hold of the differing go up against the face in PPDM and furthermore help us to locate the best fitting technique.

Mist figuring improved Internet of Things (IoT) has naturally settled noteworthy mindfulness, as the haze gadgets composed at the system fringe can manage the cost of low idleness, area cognizance yet additionally advance constant and prevalence of administrations in IoT applications. Protection safeguarding information accumulation is one of the ordinary haze registering applications, and numerous information conglomeration plans have been proposed as of late. Be that as it may, the greater part of them just helps information totally for IoT gadgets, and can't total half breed IoT gadgets' information. To address this test, R. Lu et al present a lightweight protection safeguarding information accumulation conspire, for haze processing upgraded IoT. The LPDA is recognized by applying the homomorphic Paillier encryption, Chinese Remainder Theorem, and single direction hash chain systems, yet additionally, early channel infused bogus information at the system edge. An extensive security examination shows LPDA is genuinely secure and protection improved with differential protection. Also, execution appraisals are done, which determines LPDA is really negligible.

S. Scardapane et al consider the utilization of information mining strategies in clinical settings, wherein the information to be investigated is circulated among numerous gatherings. While derivation activities could offer important clinical data, each gathering is illegal to reveal its dataset to a unified area, because of security worries of the dataset. To this end, the creator suggests an all-inclusive structure encouraging the party to execute any information mining work on depending solely on the Euclidean separation among designs, including portion strategies, unearthly grouping, etc. The issue is reworked as a decentralized network finish issue, whose arrangement doesn't require the nearness of a concentrated organizer, and full protection of the information can be guaranteed by the various systems, including irregular multiplicative updates for secure calculation. Results bolster the proposition as a proficient apparatus for grouping and characterization.

In PPDM, anonymization based methodologies have been utilized to safeguard protection. Existing writing tends to different anonymization based methodologies for safeguarding the touchy data. In any case, the anonymization based methodologies experience the ill effects of the issue of data misfortune. To limit the loss of different anonymization based grouping approaches viz. Voracious k-part and Systematic grouping calculation have been proposed. Among them, the Systematic bunching gives lesser misfortune. Furthermore, these methods use all properties during the development of an anonymized database. Therefore, the danger of presentation of touchy private data is higher. P. R. Bhaladhare et al propose two methodologies for limiting the divulgence chance and safeguarding protection by utilizing precise bunching. The principal procedure readies a disparate blend of semi identifiers and delicate qualities. The following method creates an identical change of a semi identifier and delicate property. They additionally assess approaches concentrating on data misfortune and execution time. They show the adequacy of the methodologies by contrasting them and the current calculations.

In different circulated information mining settings, spillage of the

information isn't sufficient in light of security issues. To vanquish, different security safeguarding scattered information mining practice has been recommended. V. Infant et al audit and give a study on various PPDM strategies and investigate the delegate systems for PPDM. They, for the most part, clarify the circulated protection preservation technique which makes accessible ensured arrangements utilizing the crude procedures of cryptographic methodology, for example, secure multi-party computation, mystery sharing plans, and homomorphic encryption.

As of late, safeguarding protection and guaranteeing the security of information has significant issues as private data might be uncovered by ground-breaking information mining instruments. Albeit a few systems and devices have been introduced, for the most part, execute information namelessness procedures. J. Chun et al present a novel Privacy-Preserving and Security Mining Framework (PPSF), centers around PPDM and information security. PPSF is an open-source information mining library, which offers a few calculations for (1) information secrecy, (2) protection safeguarding information mining (PPDM), and (3) security saving utility mining (PPUM). PPSF has an easy to use interface that permits running calculations and showing the outcomes, and it is a functioning undertaking with ordinary arrivals of new calculations, enhancements, and documentation.

3. PROPOSED WORK

The aim of the proposed work is to study the privacy-preserving data mining technique. Therefore a decision tree based PPDM model is proposed in this work. This section explains the strategy is used to achieve privacy-preserving data mining.

Figure 1: client server architecture

The proposed privacy-preserving data model is demonstrated in figure 1. According to the diagram the system contains a client and a server application. The client interface enabled to accept the dataset and encrypt the data in the client end. In order to keep security and privacy, the data owner encrypt the data using a private key encryption algorithm. Therefore here the AES encryption algorithm is implemented. The key reason for implementing the AES algorithm is that this algorithm is efficient and less time and memory consuming. Additionally, the algorithm provides the flexibility to use the variable size of cryptographic key. Therefore the user can use any password to encrypt the data. Figure 2 shows the cryptographic process used for encrypting the user side.

According to the given figure the user who agreed to combine and data mine with the other business parties. Produce the input part of data to the system. Additionally, a password is also provided by the user. The given password is produced to the MD5 hash generation algorithm to generate the 128-bit password. Further, the input data and the 128-bit password is produced to the AES encryption algorithm. That algorithm processes the data and generates the ciphertext. After encryption of data, the data is sent to the server for the data mining process.

Figure 2: encryption process

When the data appeared on the server, the server reorganizes the data obtained from different parties. Here all the client data is considered in a vertical partitioned format. All the aggregated data is passed on the C4.5 decision tree for generating "IF-THEN-ELSE" rules. After generating the decision rules using the C4.5 algorithm the entire rules are distributed to

all the concerning parties. At the client end, the received rules are decrypted using the previously passed cryptographic key. The benefit of this technique is that the client can only view the part of rule which is contributed by own. Additionally, the decision labels are available at all the party. Thus each party can securely obtain the secure decision rules without disclosing any part of data.

4. RESULTS ANALYSIS

The proposed privacy-preserving data mining technique is implemented in the previous section. In this section, the experimental analysis of the given model is provided. Thus different performance factors are measured and reported here.

4.1 Accuracy

The proposed PPDM technique is implemented with the C4.5 decision tree algorithm. Here the performance of the decision tree before and after encryption of the data performance in terms of accuracy of the decision tree is measured and reported in table 4.1 and figure 4.1. The accuracy is an essential factor of performance in data mining. That is measured using the ratio of correctly recognized and total samples to recognize. The following equation can help to understand.

$$\text{accuracy} = \frac{\text{total correctly classified} \times 100}{\text{total samples to classify}}$$

The table 1 shows the accuracy of both the scenarios in first before encryption the accuracy of the algorithm is computed and then after the accuracy of the algorithm on the same dataset is measured and compared in table 4.1 their graphical representation is given figure 4.1.

Dataset Size	Before Encryption	After encryption
100	78	77
300	82	78.3
500	81	77.5
800	82.8	79.6
1000	84.6	81.4
1500	84.9	81.8
2000	86.2	83.2

Figure 1 shows a line graph for accuracy, here the X-axis contains the size of datasets that are used for experiments. Additionally, the Y-axis shows the accuracy of the decision tree algorithm in percentage.

Figure 3: accuracy (%)

The red line shows the performance of the decision tree after encryption and before encryption, the performance is given using the blue line. According to the results, the encoding of data can change the application of a system, additionally, the combination can also impact of data mining application.

4.2 Encryption Time & Decryption Time

The time consumption of an algorithm is also known as time complexity. The time required to execute an algorithm with the application data is termed here as the time consumption. That is estimated using the following formula:

$$\text{Time consumption} = \text{algorithm end time} - \text{start time}$$

Dataset Size	Encryption time	Decryption time
100	103	91
300	261	255
500	302	289
800	351	328
1000	389	369
1500	447	428
2000	501	479

Table 4.2 shows the time consumption of the proposed system during the encryption and decryption of data. The measured values of time in terms of milliseconds (ms) are denoted here.

Figure 4: time consumption

The line graph representations of table 4.2 values are given in table 4.2. The X-axis of the diagram demonstrates the dataset size and Y-axis shows the time consumption of the algorithm in terms of milliseconds. According to the noticed results, the encryption and decryption approximately require a similar amount of time. But the decryption of data mostly requires less amount of time as compared to encryption.

4.3 Encryption and Decryption Memory

Memory usages or consumption is also known as the space complexity of the algorithm. The memory usages of the algorithm during the encryption and decryption of data is measured and reported in this section. JAVA technology enables us to calculate the memory usages of the processes. The following equation can be used to measure it.

$$\text{memory usage} = \text{total assigned} - \text{total free}$$

Dataset Size	Encryption memory	Decryption memory
100	16727	15253
300	16625	15277
500	16829	15529
800	17836	15376
1000	17242	15927
1500	17489	16183
2000	17352	16252

The memory usages of the algorithm are demonstrated in table 4.3. The memory is measured here in terms of kilobytes (KB).

Figure 5: memory usages

Table 4.3 values are represented in a line graph as given in figure 4.3. In this diagram, the X-axis shows the size of the dataset additionally the Y-axis shows the memory usages in terms of KB. The red line shows the decryption memory and the blue line shows the memory of encryption. According to the measured memory, the encryption takes a higher amount of memory as compared to the decryption.

5. CONCLUSION & FUTURE WORK

The proposed work is motivated to explore and investigate the privacy-preserving data mining technique. During the study, it is recognized that data mining can be used to manage the secure data modeling. In this context, a cryptographic technique is applied to the client-side on which the data owner first encrypts the data and then submits their data to the server for processing or decision rule mining. At the end of the server, only data is combined and then processed using decision tree C4.5. Finally, the recovered rule is distributed to all the clients and they can use these rules by deciphering the rules by their own private keys. On the basis of design observations and obtained outcomes of the experiments, the following facts are concluded.

- A. The encryption of data can impact on the performance of the supervised classifiers
- B. The decryption of data requires less amount of time and memory to process the data
- C. Accuracy of the decision tree also depends on other data owner's part of the information

By these concluding facts, the proposed model is extended for the following task.

- A. Improve the data model to manage the actual data utility of the different party inputs

Need to reduce the cryptographic time and memory consumption using some other alternative techniques

REFERENCES

Baby, V., N.S. Chandra. 2016. Privacy-Preserving Distributed Data Mining Techniques: A Survey, *International Journal of Computer Applications* (0975 – 8887), 143(10).

Bhaladhare, P.R., D.C. Jinwala. 2016. Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model, *Journal of Information Science and Engineering*, 32, 63-78.

Bhamare, D., P. Suryawanshi. 2019. Review on Reliable Pattern Recognition with Machine Learning Techniques, *Fuzzy Information and Engineering*, 10(3), 362–377.

Chun, J., W. Lin, P.F. Viger, L. Wu, W. Gan, Y. Djenouri, J. Zhang. 2018. PPSF: An Open-Source Privacy-Preserving and Security Mining Framework", *IEEE International Conference on Data Mining Workshops*.

Deepashri, K.S., A. Kamath, Survey on Techniques of Data Mining and its Applications, *International Journal of Emerging Research in Management & Technology* ISSN: 2278-9359, 6(2).

Kavakiotis, I., O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda. 2017. Machine Learning and Data Mining Methods in Diabetes Research, *Computational and Structural Biotechnology Journal*, 15, 104–116.

Lu, R., K. Heung, A.H. Lashkari, A.A. Ghorvani. 2017. A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT, 2169-3536, *IEEE*, 5.

Mahdavinejad, M.S., M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, A.P. Sheth. 2018. Machine learning for internet of things data analysis: a survey, *Digital Communications and Networks*, 4, 161-175.

Manikandan, V., V. Porkodi, A.S. Mohammed, M. Sivaram, 2018. Privacy Preserving Data Mining Using Threshold Based Fuzzy C-means Clustering, *ICTACT Journal on Soft Computing*, 9(1).

Mendes, R., J.P. Vilela. 2017. Privacy-Preserving Data Mining: Methods, Metrics, and Applications, 2169-3536, 5, *IEEE*.

Rajesh, N., K.A. Sujatha, A.L. Selvakumar. 2016. Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms", *International Journal of Computer Applications* (0975 – 8887), 133(7).

Scardapane, S., R. Altillio, V. Ciccarelli, A. Uncini, M. Panella. 2018. Chapter 12: Privacy-Preserving Data Mining for Distributed Medical Scenarios, *Smart Innovation, Systems and Technologies* 69, Springer International Publishing AG.

Upadhyay, S., C. Sharma, P. Sharma, P. Bharadwaj, K. R. Seeja, 2018. Privacy preserving data mining with 3-D rotation transformation, *King Saud University Journal of King Saud University – Computer and Information Sciences*, 30, 524-530.