A Study on Privacy Preserving in Big Data Mining Using Fuzzy Logic Approach

[1]M. Kiran Kumar, [2]Dr. Pankaj Kawad Kar
[1]Research Scholar, Dept. of CSE, Sri Satya Sai University of Technology and Medical Sciences. ORCID:0000-0001-9604-6799, Mail- kirann.intell@gmail.com
[2]Associate Professor,Sri Satya Sai University of Technology and Medical Sciences.
Article History: Received: April 2020; Accepted: August 2020; Published online: November 2020.

ABSTRACT: Information security is the most acclaimed issue when distributing individual information. It guarantees individual information distributing without revealing touchy information. The much well known methodology is K-Anonymity, where information is changed to comparability classes, each class having a set of K-records that are undefined from one another. Yet, a few creators have called attention to various issues with K-obscurity and have proposed procedures to counter them or stay away from them. l-variety and t-closeness are such procedures to give some examples. Our examination has shown that this load of procedures increment computational work to for all intents and purposes infeasible levels, however they increment security. A couple of procedures represent a lot of data misfortune, while accomplishing security. In this paper, we propose a novel, comprehensive methodology for accomplishing most extreme protection with no data misfortune and least overheads (as it were the important tuples are changed). We address the information security issue utilizing fluffy set methodology, an aggregate outlook change and another viewpoint of taking a gander at protection issue in information distributing. Our basically possible strategy furthermore, permits customized protection safeguarding, and is valuable for both mathematical and all out ascribes.
Keywords: Big Data Mining, Privacy Preserving, Fuzzy Logic.

Introduction: Big Data:

In relation to the outline, the comprehensive information covers the large number of structures, semi-structures and unstructured information, with a separate fee, which may be used as data. Massive information processing points on this opportunity by collecting large-scale data are not victimized by information processing schemes because of unique choices. While many items cannot be said about the amount of knowledge provided by Brobding nagian, data extraction can also be performed in real time. Huge process knowledge needs to keep the community responsive of similar programming criteria, like large data, about programs with a strong output evaluation.

Related Study:

Privacy Preserving Data Mining:

Matwin (2013) has recently carefully studied and explored the importance of privacy-preserving data management strategies. The usage of specific techniques has shown that they are able to discourage the unfair use of data mining. Any approaches indicated that every stigmatized

community could not be more concerned about generalizing data than the population as a whole. Vatsalan et al. (2013) examined the 'PRRL' methodology for the linking of datasets to organizations through the safeguarding of privacy. In order to analyze them in 15 dimensions, a taxonomy focused on PPRL methods is therefore suggested. Qi and Zong (2012) have overviewed many available privacy mining strategies based on data sharing, manipulation, mining algorithms and hidden data or regulations. With respect to the dissemination of data, a few algorithms are currently employed on centralized and distributed data for privacy security. In order to acquire joint data mining while maintaining intact private data between shared partners, Raju et al. (2009) recognized the need to incorporate or multiply protocol dependent, homomorphic encryption along with the current definition of the Digital Envelope technique. In various implementations, the methodology suggested showed significant impact.

The latest cloud services privacy protecting approach, focused on advanced cryptographic elements, was analyzed by Malina and Hajnye (2013) and Sachan et al. (2013). The solution included anonymous entry, the right to unlink and the secrecy of data transferred. Finally, this solution is used, experimental findings are collected and efficiency comparisons are carried out. In the sense of privacy preservation features and the right to preserve the same relation in other areas, Mukkamala and Ashok (2011) contrasted a series of fuzzy mapping procedures. This comparison shall be subject to: (1) the four front changes in the fuzzy function definition, (2) the introduction of seven ways of integrating different functional values of a specific data object to a single value, (3) the use of many similarity metrics to compare the initial data and mapped data.

Data mining and database discovery are two new research sectors which investigate the automatic extraction from large quantities of data of previously unknown patterns. Recent developments in data storage, data dissemination, and associated technology have opened up a new study age where current algorithms for data mining have to be re-examined from a different perspective, namely data preservation. It is well known that this news has reached a stage that risks to privacy on a regular basis are very prevalent and deserve serious thinking, without limiting the explosion of new knowledge through the Internet and other means. Privacy protection in data mining is a new path of data mining and computational database analysis that analyzes data mining algorithms for the side effects that they have on data privacy. Data mining is mostly twofold in the protection of privacy. In order for the receiver of the information not to be able to violate their personal privacy, confidential raw details such as passwords, identities, addresses etc should first be changed or removed from the initial database. Secondly, personal information that can be extracted from the database by data mining algorithms should also be omitted, since we shall mean that intelligence of this kind can similarly harm data privacy. The key aim in data mining privacy is to create algorithms to change original data in any way in order to keep private information and privacy confidential even after the operation of mining. Often widely named the "database inferencing" question is the issue where sensitive data may be extracted from published data by unauthorized users. In this article, we provide a classification and an extensive overview of the different techniques and methodologies established in the field of data protection.

The figure below provides the data mining security context.

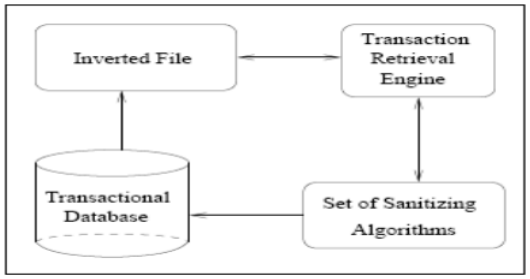

Figure 1: Privacy Preservation Classification Techniques

There are several methods to data mining security protection. We may identify them according to the following:

- Data distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first factor concerns data distribution. Some methods for centralized data have been created, whereas others appeal to a distributed data scenario. Distributed data situations can also be categorized as horizontal and vertical delivery of data. Horizontal distribution applies to these situations in which various database registers live at separate locations, thus vertical distribution refers to the cases in which all the values for different attributes are at different locations. The second component concerns the scheme for changing data. Data modifications are generally used to alter the initial values of a database to be made available to the public and to maintain high security of privacy. It is critical to provide a technique for data amendment in conjunction with an organization's privacy policy. Change methods like:

- Disturbance achieved by replacing an attribute meaning by a new value (i.e., changing a 1-value to a 0-value, or adding noise),
- block which is the substitution of an actual "?" attribute value,
- the aggregation or fusion of several values into a coarser category,
- Swapping that corresponds to the exchange in data values of each record and
- Sampling, which refers to the release of data only for a population sample?

The third dimension corresponds to the algorithm for data mining, for which the data is modified. This is not understood before, however allows the study and design of the data concealment algorithm. In our potential research agenda we have included the issue of shielding data for a blend of data mining algorithms. Different data mining algorithms are already taken into account in isolation. The key ideas for classification data mining algorithms, such as inductors of decision tree, algorithms for the mining of associations, classification algorithms, rough sets and bayesian networks have been created. The fourth factor concerns the hiding of raw data or aggregate data. Of course the difficulty in the form of rules for hiding aggregated data is greater, and most heuristics have been created for this purpose. The reduction of the volume of public knowledge leads to weaker inference guidelines for the data miner not to

enable sensitive values to be inferred. This mechanism is often called "regulatory chaos." The final, most significant dimension is the data protection strategy used to selectively modify the data. In order to gain greater usefulness with changed data, a selective alteration is needed so the integrity of data is not undermined.

The methods used for this purpose are:

- Heuristic-based methods such as adaptive modifications that only alter chosen values that mitigate the lack of efficiency instead of all possible values.
- Cryptographic methods such as protected multipart computation in which computing is safe if no entity, except its own input and results, knows something at the end of the computation and
- Approaches focused on reconstruction where the randomized data reconstruction of the initial distribution of the data.

The data manipulation results in degrading database efficiency is critical to remember. We primarily use two measures to measure the data deterioration. The first measures privacy rights and the second measures accessibility failure the second measures.

ASSOCIATION RULE MINING

Data mining is an advanced, common way of discovering the fascinating relationships between variables in broad datasets. Different measures of interest are used to analysise and display the laws contained in databases. Authors in alliance rules adopted for the detection of high-scale transaction data between items registered in supermarket point-of-sale (POS) systems.

Apriori is a classic learning algorithm for data mining. It is intended for work with transaction databases, such as collections of products purchased by clients and information on the frequency of the website. The issue is described as follows of association rule mining:

Let $\mathbf{I = \{i1, i2, ..., in\}}$ be a set of n features named items. Let $\mathbf{D = \{t1, t2, ..., tn\}}$ Be a database-listed series of transactions. Each transaction in D comprises a single ID and a subset of I elements.

An indication of the shape is known as a rule $\mathbf{X \rightarrow Y}$ where $\mathbf{X, Y \subseteq I}$. The collections of objects (in short items) X and Y are referred to as the antecedent (left hand or LHS) and the resulting (right or RHS) of the law.

The definition is shown by a brief illustration from the mobile store. The set of items is $\mathbf{I= \{JIO\ SIM, LYF, Mobile\ case, 16GB\ Memory\ Card\}}$ And there is a limited database of objects (1 is current, and 0 represents absence of item) as seen in the following Table1.2. The following is a small database. A mobile shop example rule could be

{JIO SIM, LYF} => {Mobile case}

The aim of the rule is to buy a customer mobile case if JIOSIM and LYF are purchased.

This is a really limited case. A law requires several hundred transactions in realistic implementations to be called statistics and sometimes thousands or millions of transactions can be included in data sets.

Table 1: Transactional Dataset

Transaction_ID	JIO SIM	LYF	Mobile case	16GB Memory Card
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0
6	1	0	0	0
7	0	1	1	1
8	1	1	1	1
9	0	1	0	1
10	1	1	0	0
11	1	0	0	0
12	0	0	0	1
13	1	1	1	0
14	1	0	1	0
15	1	1	1	1

PRIVACY PRESERVING ALGORITHMS

HEURISTIC-BASED TECHNIQUES

A variety of techniques for many techniques such as sorting, relationship rule discovery and clustering have been built on the grounds that selective data alteration or sanitization is an NP-hard problem and therefore heuristics may be used to deal with complexity problems [16].

Confusion over centralized data perturbation-based association rules

An optimum sanitization is a systematic evidence of an NP Hard challenge in the exploration of association rules for hiding sensitive big articles collections. The following is the basic issue that was dealt with in this work. If D is the root database, R is a collection of important association rules that can be exploited by D, and Rh is an array of rules in R. How do we convert the D_ database into the published database D_, so that all R laws, except the Rh rules, also have to be mined from D_? The heuristic proposal to modify the data was focused on data interference; in fact, the process involved changing a chosen range of 1-value to 0-values, such that the support for sensitive laws is reduced to a maximum value for the usefulness of the published database. The usefulness of this work is calculated as the amount of un-sensitive laws hidden from the side effects of the method of data alteration. The sanitation of big sensitive items to sanitizing sensitive laws is then extended to include later jobs. The methods used in the work were either to avoid the sensitive rules by covering their repeated item sets, or to reduce the trust of the sensitive rules by taking them below the user-specified threshold. Both tactics have contributed to three methods to hide delicate laws. The main feature about this respect was the ability to convert a 1-value to a 0-value and a 0-value to a 1-value in the binary database. This versatility in data alteration had the side effect that a non-frequent law might become popular apart from the secret norms of non-sensitive association. This is what we call the 'ghost law.' Since critical rules are concealed, both hidden, non-sensitive rules and frequent (ghost rules) rules are seen as less

useful than the database posted. Therefore, the heuristics used for this subsequent work should be more sensitive to utility problems as safety is not affected [19].

Confusion about centralized data blocking-based association rules

Data blocking is one of the data modifying techniques used to confuse connection rules. The blocking solution is taken by replacing a query mark with certain properties of certain data objects. Often it is more appropriate to substitute a true value with an unknown value than to put a fake value in particular implementations (e.g. medical applications). The incorporation of the new special attribute into the data collection imposes improvements in the concept of a law of association help and trust. The minimum support and the minimum trust shall be adjusted accordingly in a minimum support period and a minimum trust interval. As long as there are principles outside the center of these two ranges of help and/or trust with a critical guideline, we consider security of data to not be broken. Notice that all 1-values and 0-values should be mapped to interlayer questions on the algorithm used for rule uncertainty in such a case; otherwise it will be clear that the question marks are of origin [20].

Confusion over centralized data blocking-based classification rules

The system administrator aims to block values for the class mark inside the classification law structure. This would make it impossible for the recipient to create informative templates for the data that is not downgraded. Parsimonious downgrading is a mechanism for the formalization of the phenomenon of knowledge removal from a collection of evidence for down gradation from a safe context (called "high") to the public (called "low"), provided the presence of inference networks. In a slight decline, the possible degradation knowledge that it is not transmitted to Low is given a cost measure. The key objective to be achieved through this work is to see whether the lack of functionality due to data failure is worth the additional secrecy. In parsimonious downgrading sense, classification rules and decision trees in particular are used to analyze the possible inference channel for downgrading the results. The downgrading strategy used is to create the parametric base collection. Especially a. parameter —parameter —parameter —parameter.0—parameter —parameter 1 shall be put instead of blocked meaning. The default parameter is a likelihood of one of the attribute potential values. Pre-blocking entropy value is determined. The entropy value is calculated after blocking. In comparison with the decline in trust in the rules of the decision tree, the differential in entropy value is to determine whether the enhanced protection values value the reduction in the value of the data that Low receives. The mechanism consists of a knowledge-based policy maker, a guard to calculate the volume of data leaked and a slim downgrade to change the original downgrades decisions. The decision makers will decide the laws they should presume. The algorithm used to degrade the data determines the laws are essential to distinguish private data from those caused by incorporation of the decision tree. Data not supporting rules so discovered, along with other attributes not covered in the rule's clauses shall not be downgraded. The algorithm can determine from the remaining data which values are to be converted into missing values. This is to optimize the confusion of the law. The method of 'guard' specifies the correct level of confusion [16].

Techniques based on cryptography

In the sense of privacy protection, a variety of cryptography-based methods have been established to address the following problems. There are two or three people that wish to do a calculation on the basis of their private input, but none else will tell their results. The problem is how to do such a calculation while maintaining the input privacy. The Safe Multipart Computing (SMC) issue is called this. In specific, an SMS issue concerns computing a probabilistic function on any data, in a distributed network, with each participant holding one of the inputs, maintaining input independence, computing accuracy and revealing no more knowledge to a computing participant than the participant [18].

Conclusion: An essentially doable methodology for accomplishing most extreme protection with more data and least overheads (as just the fundamental tuples are changed) is proposed. However dimensionality decreases is proposed in before work, that isn't be fundamental in our work. The information protection issue is tended to utilizing fluffy set methodology, an absolute change in outlook and another viewpoint of taking a gander at protection issue in information distributing. The area speculation based arrangement totally disassociates the delicate qualities with the distinguishing ascribes. Our for all intents and purposes attainable area speculation technique furthermore, permits customized security conservation, and is valuable for both mathematical and absolute credits [17].

REFERENCES:

1. Abul, O., Atzori, M., Bonchi, F., & Giannotti, F. (2007, October). Hiding sensitive trajectory patterns. In Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on (pp. 693-698). IEEE.
2. Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. In Privacy-preserving data mining (pp. 11- 52). Springer US.
3. Aggarwal, C. C., & Philip, S. Y. (2008). Privacy-preserving data mining: A survey. In Handbook of database security (pp. 431-460). Springer US.
4. Aggarwal, C. C., & Yu, P. S. (2007, April). On privacy-preservation of text and sparse binary data with sketches. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 57-67). Society for Industrial and Applied Mathematics.
5. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
6. Agrawal, R., &swami A. (1993, June). Mining association rules between sets of items in large databases. In Acmsigmod record (Vol. 22, No. 2, pp. 207-216). ACM.
7. Amiri, A. (2007). Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems, 43(1), 181-191.
8. Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., &Verykios, V. (1999). Disclosure limitation of sensitive rules. In Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on (pp. 45-52). IEEE.

9. Belwal, R. C., Varshney, J., Khan, S. A., Sharma, A., & Bhattacharya, M. (2008, October). Hiding sensitive association rules efficiently by introducing new variable hiding counter. In Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on (Vol. 1, pp. 130- 134). IEEE.
10. Berberoglu, T., & Kaya, M. (2008, May). Hiding fuzzy association rules in quantitative data. In Grid and Pervasive Computing Workshops, 2008. GPC Workshops' 08. The 3rd International Conference on (pp. 387-392). IEEE.
11. Bertino, E., Fovino, I. N., &Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery, 11(2), 121-154.
12. Bononi, L., Bracuto, M., D'Angelo, G., &Donatiello, L. (2005, June). Concurrent replication of parallel and distributed simulations. In Principles of Advanced and Distributed Simulation, 2005. PADS 2005. Workshop on (pp. 234-243). IEEE.
13. Borhade, S. S., & Shinde, B. B. (2014). Privacy preserving data mining using association rule with condensation approach. International Journal of Emerging Technology and Advanced Engineering, 4(3), 292-296.
14. Brijs T., Swinnen G., Vanhoof K., and Wets G. (1999), The use of association rules for product assortment decisions: a case study, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), August 15-18, pp. 254-260. ISBN: 1-58113-143-7.
15. Brin, S., Motwani, R., Ullman, J. D., &Tsur, S. (1997, June). Dynamic itemset counting and implication rules for market basket data. In ACM SIGMOD Record (Vol. 26, No. 2, pp. 255-264). ACM.
16. K. Bhargavi. An Effective Study on Data Science Approach to Cybercrime Underground Economy Data. Journal of Engineering, Computing and Architecture.2020;p.148.
17. S. Jessica Saritha. AN EFFICIENT APPROACH TO QUERY REFORMULATION IN WEB SEARCH, International Journal of Research in Engineering and Technology. 2015;p.172.
18. K BALAKRISHNA,M NAGA SESHUDU,A SANDEEP. Providing Privacy for Numeric Range SQL Queries Using Two-Cloud Architecture. International Journal of Scientific Research and Review. 2018;p.39
19. K BALA KRISHNA, M NAGASESHUDU. An Effective Way of Processing Big Data by Using Hierarchically Distributed Data Matrix. International Journal of Research.2019;p.1628
20. Sehgal.P, Kumar.B, Sharma.M, Salameh A.A, Kumar.S, Asha.P (2022), Role of IoT In Transformation Of Marketing: A Quantitative Study Of Opportunities and Challenges, Webology, Vol. 18, no.3, pp 1-11

21. Kumar, S. (2020). Relevance of Buddhist Philosophy in Modern Management Theory. Psychology and Education, Vol. 58, no.2, pp. 2104–2111.

22. Roy, V., Shukla, P. K., Gupta, A. K., Goel, V., Shukla, P. K., & Shukla, S. (2021). Taxonomy on EEG Artifacts Removal Methods, Issues, and Healthcare Applications. Journal of Organizational and End User Computing (JOEUC), 33(1), 19-46. http://doi.org/10.4018/JOEUC.2021010102

23. Shukla Prashant Kumar, Sandhu Jasminder Kaur, Ahirwar Anamika, Ghai Deepika, MaheshwaryPriti, Shukla Piyush Kumar (2021). Multiobjective Genetic Algorithm and Convolutional Neural Network Based COVID-19 Identification in Chest X-Ray Images, Mathematical Problems in Engineering, vol. 2021, Article ID 7804540, 9 pages. https://doi.org/10.1155/2021/7804540
24. P.Padma, Vadapalli Gopi,. Detection of Cyber anomaly Using Fuzzy Neural networks. Journal of Engineering Sciences.2020;p.48.