Privacy Preserving in Data Mining with No Data Loss with a Combinational Scheme

A.Damodar, C.Rajeev, M.Srinivas Reddy

Abstract:- Large amounts of data collected by many organizations under-goes data mining for various purposes like analysis and prediction. During data mining tasks, the sensitive information may be losing its privacy. Hence, Privacyprotection or preservation is becomes major issue for the organizations. Publishing data or sharing information for mining with Privacypreservation is possible through Privacypreserve data mining technique (PPDM). Existing techniques are not able to withstand for some attacks and some suffers with data misfortune. In our paper we conventional an effective and combinational approach for security safeguarding in information mining. Our approach with can withstand from different kinds of assaults and limits data misfortune and increases data re-usability with data reconstruction capability.

Keywords - Privacypreserving, sensitive information, Data mining, K-anonymity, Randomization

I. INTRODUCTION

Preserving Privacyis a challenging task in many of data mining applications like credit card companies, health care sector and other sectors where lot of data is collected day by day. This large amount of data collected is given for data miner for various kinds of analysis and to gain some useful knowledge. Further they can predict some conclusions and act according to knowledge gained. But the data provided for miners may have private or sensitive information. In case of organizations such as credit card companies maintain lot of customer's information. This information may be provided to data miner for analysis of various things like regular users, occasional users to give offers. Can get defaulters information whether to sanction new loans or not. However, in the process of getting useful knowledge by the miner may explore / expose some sensitive information. So Privacy becomes a serious issue when data has sensitive information. Shaded part in figure 1 shows the position where exactly our proposed model works in Data mining process. Privacy preserving datamining techniques are emerged in the field of data mining to solve Privacy protection problems [1

Revised Manuscript Received on December 22, 2019.

A.Damodar, Asst. Professor, CSE., Malla Reddy Engineering College for Women, Hyderabad, India.
C.Rajeev, Asst. Professor, CSE., Malla Reddy Engineering College for Women, Hyderabad, India.
M.Srinivas Reddy, Asst. Professor, CSE., Malla Reddy Engineering College for Women, Hyderabad, India.

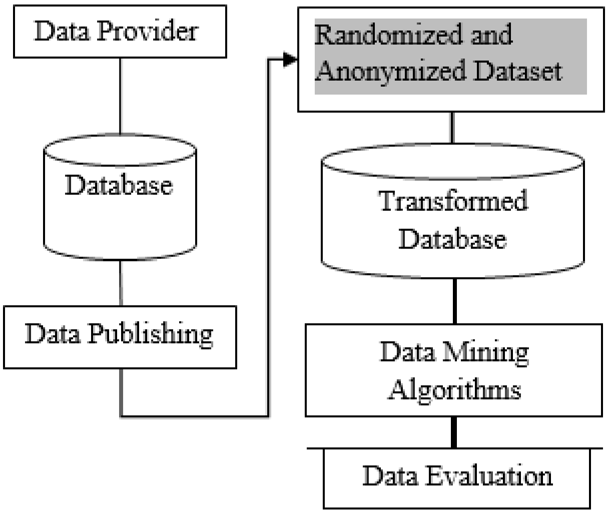

Fig1. Combinational in Privacy preserving Data Mining

Privacy preserving datamining techniques are emerged in the field of data mining to solve Privacy protection problems [1] [2]. The main motive of these techniques is to provide Privacy for sensitive information during the process of extracting useful knowledge in data mining. Into two broad areas the Privacy preserving data mining techniques are divided as data hiding and knowledge hiding. In data hiding techniques confidential or private data is completely removed or modified before sharing with a data miner. Whereas, in Knowledge hiding concentrates on concealing the delicate information that can be mined by utilizing information mining calculations [3]. A decent number of procedures have been developed for the security protection in information mining forms goes under either non-cryptographic or cryptographic. K-secrecy, I-assorted variety, t-closeness, irritation and affiliation rules goes under non-cryptographic category, whereas cryptographic methods work with multi-party computations. Power as that of cryptographic methods, even they give accurate results compared to non-cryptographic methods. But the non-cryptographic techniques suffer from information loss and to get tradeoff between information loss and Privacywe proposed an effective Privacypreservation approach. Our approach is to apply randomization on original data first and classify into more sensitive and less sensitive fields. Next we apply k-anonymization technique more sensitive tuples by leaving less sensitive tuples untouched. By this we can improve data re-usability and simultaneously reducing loss of information. This great combination makes very tough to attack on database by attackers.

Retrieval Number: D11071284S319/2020©BEIESP
DOI:10.35940/ijrte.D1107.1284S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. RELATED WORK

Completely assessment is done on stray bits of security ensuring information mining and it fundamental to indentify different ways to deal with oversee save the private or fragile data. The key goal of forming study is to get mind boggling discernment of existing security saving methodologies, and to see central focuses and deficiencies in saving private or delicate information.

A. K–obscurity Method of lucidity

Right when downsized scale information is the unconfined for an examination reason, one needs to limit exposure chance while heighten an utility of information. Sweeny well-known a k-secret method with limit a divulgence chance [4]. K - secrecy basics says that, an informational rundown is the k cloud (k 2: 1) if each record in an illuminating rundown is the ill defined from at any rate (k-l) different records inside tantamount instructive rundown. This k-uncertain quality basic is the everything considered accomplished by utilizing speculation and camouflage [5]. In principle a trademark qualities are condensed in the specific between time [6] [7]. In covering a trademark respects are supplanted or changed with some different attributes. Covering contains data catastrophe so it is the for the most part avoided. K-darken strategies inclined to partner assault [8].

B. Random disturbing

In unusual aggravation a security of the information that can be ensured by irritating a delicate information with randomization include In this paper we have chosen to utilize non-cryptographic techniques since they doesn't require high calculation of instructive records, it is an essential to preprocess a first educational list. This paper uses a system for ordinary locale to disperse a diligent data. Discrete condition is an according to an after: the (most extreme) – the (min)/n =length. A relentless attributes, n is various discrete, length is the length of discrete breaks. At a moment that a between time length is the decimal, round to a nearest entire number, an essential interval of discrete beginning from the (min), a last break is a the (most extreme). A procedure [10] [11] doesn't repeat a first data regards, and it's simply redo transport.

C. Blocking based technique

Blocking system applies to applications where we can store cloud qualities for explicit traits, when real attributes are not accessible or private [1] [2] .This method replaces I's or D's by questions ("?") in picked exchanges. Thusly, that standard won't be produced using a dataset. A target of this calculation is to cloud the given strategy of delicate standard by supplanting known attributes with obscure ones. For every touchy standard, it channels an essential database and discover outs an exchanges supporting fragile principles. We can say exchange underpins any standard when a left half of a standard pair is the subset of trademark qualities pair of an exchange and a correct hand side of a standard is same as a class property of an exchange. After that for each exchange that supports delicate principle, figuring places "?" (Unknown) values as opposed to property estimation which shows up in rule. This system proceeds until each of the a touchy standards are hidden. At long last a sanitized dataset which contains cloud qualities is discharged to open [2] [3].This technique is unquestionably not difficult to execute in any case gives data misfortune.

D. Cryptographic procedure

B. Pinkas displayed cryptographic structure [2] [4], is standard for the most part for 2 distinct reasons. Reason one is that it gives the well-depicted model to security. Another Reason is that there exist the epic toolset ofcryptographic check and systems to execute security saving information mining estimation. In any case, late work shows that this technique doesn't confirm a yield of the figuring and it's basically kill a protection spill during a period spent calculation. So it doesn't give a total course of action of security guaranteeing information mining. Precisely when in plenitude of scarcely any get-togethers are fused than this system is hard family member. In like way, it doesn't get a handle on an issue of whether an introduction of an indisputable result of data mining may break a security of independent files [15].

III. PROBLEM STATEMENT

Existing techniques for security protecting in information mining are prone to some disadvantages like data re-usability and information loss. Our research mainly focused on combining randomization technique with k-anonymity technique to Privacypreservation with increased data re-usability and reduced information loss. Consider a dataset X with attributes a1, a2….an. As for our proposed method to achieve privacy, let us apply transitional probability matrix for randomization in dataset X. Let us consider high sensitive dataset Y after randomization. Now this dataset Y undergoes k-anonymization process, which minimizes data loss. It is observed that after applying randomization &generalizations we cannot distinguish the data records of X from Y more accurately or precisely.

Our research work mainly focuses on:

a. **Data usability**: Data usability influences the exactness of information mining errands. So our goal is to dispose of or at least minimize Privacybreaches and to increase the data usability.

b. **Privacy:** We proposed a combinational methodology using randomization and k – anonymity to provide double protection for sensitive data.

c. **Data Loss:** Data or information loss should be minimized, so in our effective approach we used k– anonymity which minimizes information loss.

IV. PROPOSED METHODOLOGY

The proposed framework utilizes the mix of randomization and k-anonymization procedures. This procedure gives:

- Protection to delicate information with immaterial loss of data.
- Increases information utility and
- Data age.

Our proposed method is commonly disconnected into two sections. To some degree - I use quality transitional likelihood structure technique to

get randomization on a dataset and fairly - II k - secrecy is performed on consequence of part – I i.e., on randomized dataset. Part I:

Information: Original dataset X, Transitional likelihood cross section P, I *j size mapping cross section M which is among X and P. Yield: Converted table C. System:

a) Select semi identifier. Besides, the key traits of dubious quality from tableX.

b) Remove/Suppress/alter key properties.

c) Generate transitional likelihood cross section P with size j*j aimlessly.

d) Generatemapping cross section M aimlessly.

e) Rendering the mappinggrid M doleout every P (Pl,P2... Pj) to X (X1,X2 ...Xj).

f) With regard to a most raised area of P respect, modify the portion of T. In the event that most significant zone is beginning at now utilized, by then go for the going with a higher area of P. On the off chance that estimation of P of at any rate two zone was same as it will pick a left hand side worth.

g) Re-substitute in table.

h) RecombineX cross area.

Fairly I shows that first semi identifier, delicate trademark and key property are investigated tableX. Removed or covered delicate information of people can be indentify particularly by utilizing Key attributes. After that unusual time of likelihood structure P and mapping gridM. Eventually parts ofX are improved from the most raised estimation of domain ofP. If there should rise an occasion of most basic locale is beginning at now utilized by at that point select next a higher area ofP. In the event that in any occasion two zone of P gets same attributes, by then pick the left hand side worth. In tableX, All attributes are re-substitutable. After finish of part – I has changed over tableC on which k–anonymization method is applied. Ordinary k-namelessness strategy by and large considers all tuples as in like manner delicate, so all tuples has anonymized and prompts data hardship. To defeat this issue, we organized dataset into high delicate sort and low touchy sort before applyinganonymization system.

Information:

Result of Part-I i.e., changed over tableC is given as data, Anonymization parameter k.

Yield:

Last deduced tableD.

System:

a. Select the tableC.

b. Categorize atouchy trademark qualities into2 class high(H) and low(L).

c. For each tuple whose straightforward attributes have a spot with class H-Move those tuple into tableD I and apply surmise on semiascribes to anonymized it.

d. For eachtuple whose touchyqualities have aplace with class L-Move these tuples into tableD2 and don't anonymized it.

e. Append linesof tableDI and tableD2 and get last interpreted tableD. D = DI+D2.

V. CONCLUSION

Even though lot of research has done in Privacypreservation in data mining and proposed many approaches and techniques have some pitfalls. These pitfalls should overcome and should protect the Privacyof an individual. In terms of complexity and efficiency non – cryptographic techniques like anonymization and cryptographic are considered to be better when compared to cryptographic techniques. But anonymization techniques suffer with high information loss and prone to attacks. Randomization techniques do not facilitate the data utility. The proposed hybrid and effective approach will provide high Privacywith significantly less information loss and can be reconstructed easily.

REFERENCES

1. Jian Wang, Yong Cheng Lou, Yen Zha Jiajin Le, "A Survey on PrivacyPreserving Data Mining", International Workshop on Database Technology and Application pp. III - 1 1 4, 2009.
2. Agarwal, R. and Shrikant, R. "Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Data. pp. 439 - 450, 2000.
3. V.S Verykios, A.K Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, "Association Rule Hiding", IEEE Transaction Knowledge and Data Engineering, 1 6 (4): 434 - 447, 2004.
4. L. Sweeny, "K-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowlwdge based Syatem, pp. 557 - 570,2002.
5. S. Vijayrani, A. Tamilarasi, M. Sampoorna, "Analysis of PrivacyPreserving k-anonymity Methods and Techniques", Proceeding of the International Conference on Communication and Computational Intelligence, pp. 540 - 545, December 2010.
6. K. Wang, P.S. Yu and S. Chakraborty, "Bottom Up Generalization: A Data Mining Solution to PrivacyProtection", In International Conference on Data Mining, pp. 249 - 256, 2004.
7. B. Fung, K. Wang, P. Yu "Top Down Specialization" For International Conference on Data Engineering (lCDE' 05), pp. 205 - 216.
8. E. Poovamal, M. Ponnavaikkp, "Task Independent PrivacyPreserving Data Mining on Medical Data Set", International Conference on Advance Computing, Control and Telecommunication Technologies, pp. 815 - 818, 2009.
9. X. Zhang, H. Bi, " Research on PrivacyPreserving Classification Data Mining on Random Peturbation", International Conference on Information Networking and Automation (ICINA), pp. 173 - 178,2010.
10. H. Karagupta, S.Datta, Q. Wang and K. Sivakumar, "Random Data Peturbation Techniques and PrivacyPreserving Data Mining", IEEE International Conference on Data Mining 2003.
11. Dr. Saravanakumar, Dr. S.P.Anandaraj, Dr.P.Manikandan, "Spot Detection For Morphological Convolution Using DNA Microarray" , International Journal of Pure and Applied Mathematics, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version), SCIE 2013 Impact Factor= 7.19, Scopus SJR=0.322, H-index=28,G-Index=48, Vol.118, Issue 14,pp.1-7,Feb 2018.
12. K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Peturbation", Proceedings of the Fifth International Conference of Data Mining (lCDM' 05), pp. 582 - 589, 2005.
13. Anita A. Parmar, Udai Pratap Rao," Blocking Based Approach for Classification Rule Hiding to Preserve The Privacyin Database", International Symposium on Computer Science and Society (ISCCS), pp. 323 - 326, 2011.
14. J. Feiliu, J. Luo and J. Zhexue Huang, "Rating: PrivacyPreservation for Multiple Attributes with Different Sensitivity Requirements:, International Conference on Data Mining, pp. 666 - 670, 2011.
15. Vassilios S. Verykios, Elisa Bertino, "State of the art in PrivacyPreserving Data Mining", Proceeding of Special Interest Group on Management of Data (SIGMOD) Record, Vol. 33, No.1, pp. 50 - 57, 2004.
16. Y. Lindell, B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology 5(3), 2000.

Retrieval Number: D11071284S319/2020©BEIESP
DOI:10.35940/ijrte.D1107.1284S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication