International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 8, August-2018

# Prevention of confidential and sensitive information leakage in data mining

Kishore Mishra, Associate Prof. Brij Kishore
Computer Science & Engineering
Apex Institute of Engineering And Technology, Jaipur, India

*Abstract*—**Extensive measure of data is delivered in electronic form by different legislative and non-legislative organizations. Information related to specific individual needs to be protected, sothat it may not harm the privacy.It is observed that various techniques has been proposed earlier times as a mechanism for protecting privacy in publishingof sensitive data, but the mechanisms are inadequate to safeguard the secrecyproblems.K-anonymity has been used as successful techniqueitexpress this methodologyachievesenhanced over the distinct l-diversity measure, probabilistic l-diversity measureand k-anonymity through t-closeness measure since only rarer partitioning must be done for a robustsecrecyrequirement**.

***Keywords—Privacy preserving techniques, k-anonymity, Anonymization, data mining.***

## I. INTRODUCTION

With the rapid development of mobile Internet and the popularity of smart mobile terminals, more and more people use smart phones, tablet PCs and other mobile devices to access the Internet for social activities.[1] The information being gathered incorporates private or delicate information. An ever increasing number of information mining systems are appealing now a days. The data mining techniques are used to extract the hidden knowledge from huge data collections in the form of trends, models and patterns.[2] Privacy is one of basic human rights along with the emergence of dignity, right and value, and need to be respected and protected by people in their social life and communications. Privacy including people's right to live without being disturbed and the dominant control over personal information. How to secure these online private data is the problem encountered by every country.[3]If the user initiates the query continuously in a period of time, and the attacker continuous monitoring, it may be inferred that the user's trajectory or the location of the next moment and other important information, to the user's personal and property security poses a serious threat. Therefore, how to effectively protect the sensitive data has become an important research topic.[1]K-anonymity is a critical technique for security protection while discharging miniaturized scale information. So in today's environment, a data holder, such as a medical institute, public health agency, or banks, share person records in such a way that the published information remains practically useful but the identity of the individuals also determined. So to prevent this released information must follow k-anonymity.[4]

## II. PRIVACY PRESERVING TECHNIQUES

A brief overview about specific elementary and current advance approaches such as Randomization method, Anonymization method, Encryption method. There are various methodologies which have been assumed for privacy protective data mining.

### *A. Randomization Method*

The randomization method gives a powerful yet straightforward processfor keeping the client protection from learning sensitve data, which can viably executed for security keeping up data mining. In this method the additional data added to a given record is independent of the behavior of other original data records. When the randomization technique is done, the information accumulation process comprises of two phases.In the first phase, the informationsuppliers to randomize their information and transmit the randomized information to the information receiver. In the second stage, the data recipient assesses the first appropriation of the data by utilizing a distribution reconstruction algorithm.The model of randomization strategy is appeared in the Figure 1 Symbolic randomization methods contain random-noise-based agitation and randomized response scheme. The author Agrawal and Srikant describe a scheme for privacy preserving data mining using random perturbation and deliberated how restructured disseminations might be utilized for data mining.

 

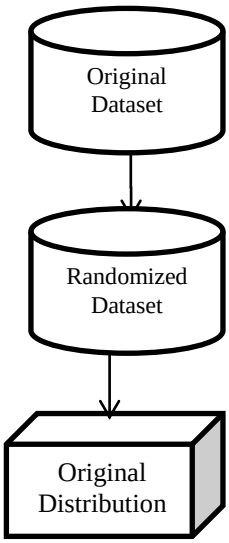


Figure 1: Randomized Model

*B. Anonymization Method*

Anonymization method objectives at building the individual record be indistint among a group record by using procedures of speculation and concealment. Various propelled strategies have been prescribed, like p-sensitive k-anonymity, M-invariance, Personalized anonymity, (a, k) –anonymity, l-diversity, t-closeness and so on. The anonymization technique can ensure that the converted data is correct, but it also results in info loss in severallevel.

For instance, various common data characteristics such as race, birth, sex, and zip are accessible in public records such as voter list and a particular data set for instance medical data, they can be used to accomplish the identity of the equivalent individual with high probability by linking process, as is shown in the figure 2.

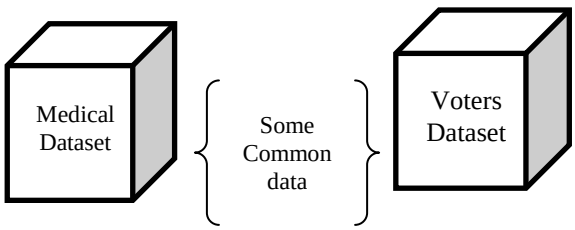


Figure 2: Anonymization Model

The k-anonymity methods mainly deliberate on a worldwide approach that utilizes the similarquantity of protection for all individual data, without providing for their solid needs. The result might offer deficient security to a subset of individuals, while applying extremesecrecy control to another subset.

*C. Encryption Method For Distributed Privacy Preserving Data Mining*

Most privacy preserving distributed data mining algorithms are produced to unveil nothing other than the last result. The privacypreserving association rule mining problem over horizontally partitioned data these methods integrate cryptographic techniques to minimize the data collective, while count immaterial overhead to the mining errand.

A Naive Bayes classifier for classification protection on vertically isolated information and the technique for clustering over vertically apportioned information. Every one of these methods are almost in light of the exceptional encryption convention known as Secure Multiparty Computation (SMC) innovation. [5]

## III. LITERATURE SURVEY

M. Prakash and G. Singaravel [2015],utilized Top-Down Greedy algorithm in which it partitions the high-dimensional space into areas and by the area's representation itencodes data focuses in one area.

Here four privacy measures are studied-

- Distinct l-diversity measure with default value of l = 5.
- Probabilistic l-diversity measure with default estimation of l = 5.
- k-anonymity with t-closeness measure with default estimation of k = 5 and t = 0.15.
- Proposed personalized anonymization approach with default estimation of k = 5, n = 1000 and t = 0.15.[2]

Vijay Sharma[2014]provides a surety that released data is in any case k-anonymous. Numerous methods have been suggested to realize k-anonymity for the specified dataset. It categories these methods into four main fields based on the standard these are based and methods they are applying to accomplish k-anonymous data.Four main approaches to the solution of k-anonymity problem for anonymization of data have been discussed. These methodologies have been diplayed in a basic and easy to understand language, with examples. All these categories include most of the algorithms developed so far for accomplishing k-anonymity[6]

Fei Liu et al [2013]recommend a new k-anonymity algorithm for sensitive characteristics. He divide sensitive attributes values into highly sensitive ones and lowly sensitive ones. Tuples are sorted according to amount of highly sensitive values first and then distributed to best equivalence classes one by one. We destroy association among sensitive attributes values to avoid attack. He introduce information entropy to evaluate diversity of equivalence classes.[7]

Mohammad Reza Zare Mirakabad et .al [2008] recommend procedures to discover several questions about k-anonymity of data. Such questions are, for example, "is my informationadequatelyunknown?", "which info, if accessible from an outside source, threatens the anonymity of my data?".The methodology that they propose affect two properties of k-anonymity that they express as two lemmas. The principal lemma is a monotonicity property that empowers to adapt the A-priori algorithm for k-anonymity. The following lemma is a determinism property that empowers to devise an effective algorithm for δ-suppression.[8]

Xiangwen Liu et .al [2015] propose a personalized extended (alpha, k)-anonymitymodel for the purpose of personalized privacy preservationrequirements in Privacy Preservation Data Publishing technology.Our model combines sensitive attribute value-orientedprivacy preservation method with individual-oriented method,and unifies the privacy protection requirement of above twomethods with Privacy Preservation Level.The experimental results show that our model can providestronger privacy protection with not much time cost. d. Experimental results show that the personalized extended (α, k)-anonymity model can provide stronger privacy protection efficiently.[9]

JordiSoria-Comas et .al[2013] approach combines *k*-anonymity and *ε*-differentialprivacy to reap the best of each approach for anonymizeddata publishing: namely, the reasonably low informationloss incurred by *k*-anonymity and its lack of assumptionson data uses, and the robust confidentialitypromises offeredby *ε*-differential privacy. Theyusage a new defined insensitivemicroaggregation to achieve a *k*-anonymous data setby considering all features as quasi-identifiers; then take the *k*-anonymous microaggregated data set as aninput to which uncertainty is added in order to reach*ε*-differential privacy which show that our combined approach reduces information loss by several orders magnitude[10]

Shyue-Liang Wang et .al[2011]suggested two procedures, *Sensitive Transaction Neighbors*(*STN*) and *Gray Sort Clustering* (*GSC*), by addition/deletion of*QI*items and adding *SI* items to realize*sensitive k-anonymity* ontransactional data. Extensive numerical researches wereassumed to determine the features of the suggestedconcept and approaches.It is different from *k-anonymity* on transaction in that transactionsmay contain both sensitive items and quasi-identifying items.[11]

Tanashri Karle et .al [2017] main focus of the study is Privacy Preservation using Anonymization Technique and a detailed study two Anonymization Algorithms are explained – Datafly Algorithm and Mondrian Algorithm. Datafly algorithm is more suitable for synthetic dataset while Mondrian algorithm is more suitable for real dataset. Datafly Algorithm performs better when dataset is We have done a detailed study of these two algorithms and achieved a detailed comparison of these two algorithms based on thirteen parameters.[12]

## IV. COMPARATIVE ANALYSIS OF PRIVACY PRESERVING TECHNIQUES

| Method | Advantage | Disadvantage | Approach |
|---|---|---|---|
| k-anonimity | It reduces the granularity of data representation.<br><br>This granularity is decreased adequately that any given record maps onto in any event k different records in the data. | The technique is susceptible to many kinds of attacks specially when background knowledge is available to the attacker.<br><br>The foe can utilize a relationship between at least one identifier qualities with the sensitive attribute keeping in mind the end goal to limit conceivable estimations of the sensitive field further. | k-anonymous method |
| Randomization | Data is changed by adding noise to the original data.<br><br>Identification of data directly is not possible.<br>The first record esteems can't be effectively speculated from the distorted data.<br><br>It is relatively simple, and does not require knowledge of the distribution of other records in the data | The method on its own is weak and does not offer complete reliability, hence it is used in combination with other algorithms.<br><br>The quality of data is disturbed and the procedure is irreversible.<br><br>Reproductions prompts the spillage of Privacy, which identifies with the conceivable dangers. | Added substance Perturbation<br>Bother by irregular projection system |
| Encryption | The technique bunches the information into different classes and the encryption depends on the key qualities produced inside each class.<br><br>Since the key is not a constant private or public key, the method provides a greater amount of protection. | It involves complex mathematical computations. | Integer partitioning based encryption |
| Cryptography | Separate parties can jointly compute any function of their inputs, without revealing any other information.<br><br>It conceals all data aside from the assigned yield of the function | There may exit Corrupted gatherings, who pick their sources of info freely of the legitimate gatherings' information sources.<br><br>This property is pivotal in a fixed auction. | Oblivious exchange |

## V. CONCLUSION

It is examined that the effectiveness with respect to various quasi identifier sizes of customizedanonymization approach is sufficiently quick to be utilized. The efficiency factor is figured by fluctuating k and l valuesfor the proposed approach, it also comparatively enhanced over the distinct l-diversity measure, probabilistic l-diversityquantity and k-anonymity with t-closeness quantity.

REFERENCES

[1] Wen He,"Research on LBS Privacy Protection Technology in Mobile Social Networks", 2017.

[2] M. Prakash and G. Singaravel, "An approach for prevention of privacy breach and information leakage in sensitive data mining",Computers and Electrical Engineering,2015.

[3] Huancheng Liu, Xiaolong Liu," The Protection of the Privacy Right in Electronic Commerce", 2012.

[4] B.B.Patil and A.J.Patankar, "Multidimensional k-anonymity for Protecting Privacy using Nearest Neighborhood Strategy", IEEE International Conference on Computational Intelligence and Computing Research, 2013.

[5] Ms. Dhanalakshmi.M,Mrs.Siva Sankari.E, "Privacy Preserving Data Mining Techniques-Survey", 2014.

[6] Vijay Sharma, "Methods for Privacy Protection Using K-Anonymity",International Conference on Reliability, Optimization and Information Technology,India, Feb 6-8 2014.

[7] Fei Liu et .al, "A New k-anonymity Algorithm towards Multiple Sensitive Attributes", IEEE 12th International Conference on Computer and Information Technology, 2012.

[8] Mohammad Reza Zare Mirakabad1 et.al, "Towards a Privacy Diagnosis Centre: Measuring *k*-anonymity", International Symposium on Computer Science and its Applications, 2008.

[9] Xiangwen Liu et.al, "A Personalized Extended (_, k)-Anonymity Model", Third International Conference on Advanced Cloud and Big Data,2015.

[10] Jordi Soria-Comas et.al, "Improving the Utility of Differentially Private Data Releases via k-Anonymity", 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications,2013.

[11] Shyue-Liang Wang et.al, "K-anonymity on Sensitive Transaction Items", IEEE International Conference on Granular Computing, 2011.

[12] Tanashri Karle,Prof. DeepaliVora "Privacy Preservation In BigData using Anonymization Techniques", InternationalConference on Data Management, Analytics and Innovation (ICDMAI), 2017.