International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 6 4244 – 4248

Maximizing Data Utility by using Data Anonymization Technique

Ms.Padma L.Gaikwad[1]
[1] Department of Computer Engineering, SVIT,
Chincholi-27,
Nashik , India
padmagaikwad14@email.com

Prof.M.M.Naoghare[2]
[2] Department of Computer Engineering, SVIT,
Chincholi-27,
Nashik , India
manisha.naoghare@gmail.com

Abstract— Data Anonymization is one of the techniques for achieving the database privacy. In this method of data anonymization the facility of hiding the identification factor from the other users. Hence the feature of this method is used for the modification purpose only and that will remove the identities of person and appears table in same way. It will help for maintaining the risk factor. This approach can be increased data utility tradeoff in an organization. Scalability and privacy risk are the main factors regarding any database over an organization. Here this are two factor can be help to maximizing the data utility as well as minimizing the risk by using differential privacy preserving method. At the time of released data differential privacy preserving mechanism support for individual data hiding , by adding the noise and disclose for the secondary purpose.

Keywords- *Data Anonymization, Differential, Privacy, scalability,diversity.*

I. INTRODUCTION

Many organization works on the real time data and they want to personal information for the investigation purpose. In health care system the patient need to fill all the necessary personal information. In government sector the personal information includes all the necessary individual data regarding that person. Such organization can use the collection of large dataset for the secondary purpose by hiding the identities. To maintain a database privacy and provide security over the database here the data anonymization technique used under different suitable mechanism and algorithms. Because the anonymization method can only hide the one or two identities from the table ,hence here the differential privacy preserving mechanism help us to provide mathematical bound for protecting the information and once the database bound within a range there are minimum chances to miss the data from the dataset. Before data released apply the necessary transformation for achieving the privacy and security over the database community. Data disclosure method is more advantageous in an organization for achieving the data privacy and data security. Privacy for the database is becoming a huge problem in many areas such as government, hospitals; many companies etc. Data Anonymization is a one type of technique that is used for conversion of clear text into a non-human readable form. It is used to enable the publication of detail information. Basically data anonymization provides the privacy guarantee for the sensitive data against the various attacks over the database community. To achieve privacy guarantee there are two different techniques such as K-anonymization and l-diversity. K-anonymization is one of the technique which includes the hiding of identities and it is more accurate technology for the data anonymization. There have been no evaluations of the actual re-identification probability of k-Anonymized data sets. In k-anonymization each record is distinguishable from k-1 records with respect to certain identifying records. One of the limitations of k-anonymization can overcome the l-diversity. K-anonymization does not provide the privacy guarantees against the attacker using background knowledge. L-diversity is a more powerful technique that can overcome the weaknesses of k-anonymity. K-anonymity is not always effective in preventing the sensitive data of the dataset. The technique of l-diversity is used to maintain the group of sensitive attributes for protecting the data against the background attackers. Characteristics of l-diversity are to treats all values of attribute in a similar way irrespective of distribution in the data. L-diversity is achieved to difficult for sensitive data. It gives different degree of sensitivity. L-diversity does not consider overall distribution of sensitive values of the record set because of equivalence classes on quasi-identifier. It does not consider semantics of sensitive data. The t-closeness is one of the techniques ensuring the distance between the distribution of sensitive attributes in a class of records and the global distribution .In t-closeness the distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database.

There are different techniques of an anonymization such as:

1. Data Suppression:-In this technique the information is removed from the data. For example the gender field can be removed from the dataset.

2. Data Generalization:-In this techniques the information is coarsened into set or range .For example age of the person can be display in range form

.3.Data Perturbation:-In this technique noise is added directly between the entities. For example the pin code of city can be display in addition of noise form.

The differential privacy preserving algorithm provides both scalability and privacy risk by using various polynomial algorithms. Differential privacy provides an interesting and rigorous framework around publishing data. Differential privacy provide to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. Privacy is important when the contents of a message are at issue and whereas anonymity is important when the identity of the author of a message is at issue. The role of privacy preserving algorithm which prevent the leakage of specific information about person. Sensitive input data is randomized, aggregated, Anonymized and generally contorted to remove any concrete implication about its original form.

II. LITERATURE SURVEY

Data Anonymization is a very powerful technique for protecting the data from the various attackers and risk Although data disclosure is advantageous for many reasons such as research purposes, it may incur some risk due to security breaches [1]. Disclosure of data having the aim to limit the amount or nature of specific information that leaks out of a data set.

ARUBA and SABRI are the two superior schemes for performing the data Anonymization [1]. ARUBA is A risk-utility base algorithm [2]. ARUBA scheme is proposed for finding out the tradeoff between data utility and data privacy on the basis of algorithm. A risk-utility base algorithm determines a personalized optimum data transformation is based on the predefined risk and data utility models. This algorithm deals with the micro data on a record by record basis and identifies. The optimal set of transformation that will apply to minimize the risk and maintain the data utility above the certain threshold value. But there is one of the issue regarding risk-utility algorithm, does not provide the scalability and theoretical foundation for data privacy guarantee. ARUBA does not elaborate more on the different risk and utility models on the performance of different algorithm. SABRI proposed a realization of t-closeness.

SABRI is a Sensitive Attribute Bucketization and Redistribution framework for t-closeness [16]. SABRI is used for t-closeness and it adopts the information loss measures for each equivalence classes (EC) of released records individually. Bucketization partitions a table of records into buckets of similar sensitive attribute values in a greedy method. The sensitive attribute bucketization is fails to provide the theoretical foundations for privacy guarantees and data efficiency.

K-anonymity has popularly used for data anonymization [4]. It is an effective way to Anonymized micro data. In a k-Anonymized dataset, each record is indistinguishable from at least k - 1 other record with respect to certain "identifying" attributes. There are two common methods for achieving k-anonymity for some value of k. K-anonymity does not provide an efficient investigation for the multiple queries.

Following are the limitations of k-anonymization method:

1. It does not hide whether a given individual is in the dataset.
2. It is possible to be reveals individual sensitive information.
3. It does not protect against attack based on the background knowledge of dataset.
4. K-anonymity requires special method for dataset is anonymized and published data more than once.
5. The K-anonymity problem is NP-Hard even when the attribute values are ternparary and only at that time suppuration method is allowed.

t-closeness is used to find the distance between two distributions and the distance should not be more than a threshold value t. K-anonymity can protect identity disclosure while l-diversity and t-closeness can help to protect attribute disclosure.

A top-down specialization algorithm is developed by Fung et al. [17] that iteratively specializes the data by taking into account both data utility and privacy constraints. Both the approaches consider for classifying the quality of data utility. However, to preserve classification quality, they measure privacy as an uniquely individual can be identified by collapsing every subset of records into one record.

III. PROPOSED SYSTEM

In this proposed system the main focus on the problem of release statistical information about a dataset without compromising the privacy of any individual. Here the system can handle data scalability and data privacy. There are many of the techniques available that can breach the information easily over the database community .The differential privacy preserving based algorithms can provide the personalized anonymization with the help different privacy preserving algorithm. For data security an organization applies a set of transformation rules on the database before the use of data for secondary purpose. The database community contains the sensitive data as well as the quasi-identifier (QI's).The l-diversity and t-closeness can apply set of rules on different attribute such as sensitive data and quasi-identifier separately. The proposed system basically works on such a type of attribute to achieve the data privacy and also maximize the data utility.

From the literature survey the existing system is not secured one. In this system an attacker easily breach the information, because there is no mathematical bound used only the data can hide from the unauthorized user.

For achieving data privacy an exponential mechanism apply on the released data by the generalization. Generalized given attributes in an exponential manner helps for preventing data from unauthorized person, because all records are corresponds to each other. Here the utility function is considered as a record independent of the database. By using the differential privacy algorithm, maximize the scope of data utility of the data disclosure while maintain the risk below the certain acceptable threshold value. Polynomial time algorithm can reduce the number of function. An informal and formal model used for the purpose of risk-utility function under the threshold formulation .Informal model can distinguish between the feasible points and the infeasible points. The feasible points show the shaded region that belong the risk-utility tradeoff mechanism. Formal model use the concept of Value Generation Hierarchies (VGH's), that means the information is completely arranged in a hierarchical manner or in a generalized form. The very desirable property of supermodularity for the data anonymization is to minimize or maximized data privacy in a polynomial time. Polynomial time shows the result of acceptance as well as rejection of the record. Polynomial time helps to reduce the number of function. The min and max function for the privacy are used under the polynomial time algorithm. It will samples the given points according to given function. The reduction of function used for classifying the problem such as minimization or maximization problem.

IV. DIFFERENTIAL PRIVACY PRESERVING MECHANISM

Differential privacy preserving mechanism aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying or loosing its records. Differential privacy provides privacy preserving algorithm used for data disclosure. Disclosing the minimum amount of information or no information at all is try to protect the privacy of individual to whom data pertains[1].The differential privacy preserving algorithm provide a personalized anonymization on individual data items based on the specific risk tolerance of that data. Differential privacy mechanism can perform the masking operation on individual data, and it allows accurate percentages and trading. An approximation algorithm is deals with hardness under some condition to produce data transformation within constant guarantees of the optimum solution. For achieving differential privacy use the Laplace distribution to add noise probably to add noise in smallest amount required to preserve privacy.

$$f: D \rightarrow R^d$$

$$K(f, D) = f(D) + [Noise]\, d$$

The multiplicative factor used in the guarantee of scalable information for higher or lower guarantees of privacy. The noise is depending on the factor f and ε, not on the database. Another modified variant of the formulation is a polynomial time algorithm is used for data transformation. The polynomial time is a one type solvable algorithm and it will refers to time taken required for a computer to solve a problem, where this time is a simple polynomial function of the input. For NP-hard problem, there are polynomial algorithms used to solve all problems in NP-algorithm. Polynomial time algorithm can reduce the number of function that will maximize the utility of data. By using polynomial time algorithm, it refers to time taken to complete a task for calculating the time taken for data anonymization. Approximation algorithm work on the smallest value of threshold formulation, over the convex set of optimization. The purpose of approximation algorithm is used for solve linear programming and it is easier optimization than the other algorithm. Threshold value is a minimum or maximum value which serves as a benchmark for comparison or guidance and any breach of which may call for a complete review of the situation or the redesign of a system.

Differential privacy provides a mathematical way to model and bound the information gain when an individual is added or removed to or from a dataset D. It is natural way the privacy degrades when multiple operations are performed on the same set of information and since more information is exposed. But the privacy degrades in a well control manner. A randomized algorithm satisfy the (ε, δ)-differential privacy if,

$$Pr[A(D) \varepsilon B] \leq e^{\varepsilon} Pr[A(D') \varepsilon B] + \delta$$

For any two data sets D and D' that differ by at most one record and any subset of outputs B subset Range (A). Differential privacy bound the information gain when an individual is added or removed to or from a dataset. It will give

the support for query and requiring that the released data have noise added to ensure that the information for any individual can be sufficiently hidden from the user. It is used for protection purpose

Differential privacy ensures for the limited amount of additional risk is incurred by participating in the socially beneficial databases. The removal or addition of any record in the database that does not change the outcome of any analysis by much. That means it ensure the presences of an individual is protected against the attacker's.

Differential privacy preserving algorithm work on the basis of sensitivity function.

$$f: D \rightarrow R^d$$

$$\Delta f = \max \| f(x) - f(x') \|_1$$

For all x and x' differing in at most one element.It captures how great a difference must be hidden by the additive noise.

A key technique of randomized rounding of linear relaxations for approximation algorithm is used to rounding a fractional solution x to linear programming relaxation of a problem into an integral solution. An approximation algorithm maximizes the utility within a constant factor. An approximation algorithm use the Lovasz extension and randomized rounding of a vector extension for finding out the maximum utility. Lovasz extension shows that maximizing a linear function with non-negative coefficients.

Convex optimization is one type of techniques which is used in a wide range of disciplines such as many automatic control system, communication and networks, data analysis. Convex optimization is a straightforward approach was design for the linear programming. It can perform easier optimization than the other type of optimization. Differential privacy preserving algorithms apply a set of convex functions over a convex set. Convex optimization can be solved globally with similar complexity as linear programming. Many problems can be solved via convex optimization. In data privacy whenever the risk threshold is small, then the convex optimization is used in an approximation algorithm. Threshold value is used for comparison or guidance and any breach of information which may call. It is used for packing integer programs by employing the methods of randomized rounding technique by combining with number of alterations.

Steps of Approximation Algorithm:

1. Input: record a, real numbers.
2. Output: Generalization of a.
3. Define lower and upper bound real values for minimum and maximum function.
4. Execute min () and max () function by using for loop by using till the upper bound.
5. Solve the maximization problem over a convex set. $M = \max u^{a}(x)$
6. Apply randomized rounding extension method over the optimal solution corresponding element a^{+}.
7. Return maximum utility.

 ISSN: 2321-8169
4244 – 4248

An Approximation algorithm maximizes the data utility and maintaining risk below certain acceptable threshold value. It can give the guarantees to be close to an optimal solution. It runs in a polynomial time and obtains a good bound on the optimal solution. Randomized rounding method gives an o (log n) approximation.

V. SYSTEM ARCHITECTURE

The system architecture shows the complete view of the system. For anonymization of data the system first take the original database and classify that data according to different attribute such as sensitive attribute and quasi-identifier. Existing system work on the quasi-identifier and there are many chances to lose sensitive data. Hence in proposed system we focus on the sensitive data. Basically aim of this system is to protect data from the unauthorized person means to minimizes the risk and also we have to maximize the utility means use the available data for the secondary purpose. For such a goal of system the differential privacy preserving mechanism is used under the different algorithm and assumption. By using the formula of differential privacy preserving algorithm satisfy differential privacy over the sensitive data. The satisfaction of differential privacy is based on the basis of noise added into the original data.

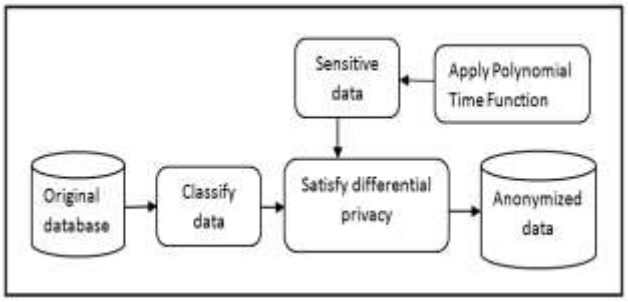

Figure 1. System architecture

The admin of the system is the only authorized person knows the formula of noise adding into the data or how much the information having mathematical bound. The differential privacy for sensitive data is work on the basis of polynomial time function. Polynomial time algorithm runs in time polynomial in the length of the input. After handling such a function the system can produced Anonymized data that will be used for the secondary purpose and investigation purpose. And the Anonymized data is fully having the mathematical bound hence no one can breach the information easily.

Such a system can be used in many sectors that will actually work on the real database, such as in banking, government, school, hospital etc. Here this system is work for the hospitals for protecting data. In any hospital first take the patient name and other essential information at the time of admit in a hospital. The registration form contain the combination of different attribute such as patient name, address, mobile number, email address, birth date, type of disease. After filling the registration form the patient can have the user name and password. System admin can provide the privacy over the original database. No one can see the sensitive information of patient and doctors also need a key for checking the details of patient.

VI. RESULT ANALYSIS

The system is used for the hospital data protection from the attackers. At the time of registration there are two different domains are used for the registration such as personal domain and public domain. Personal domain contains the patient registration and public domain perform the insurance as well as doctor domain. Form the personal domain the system can generate the graph of disease. This graph will be useful for the secondary purpose for investigation of disease.

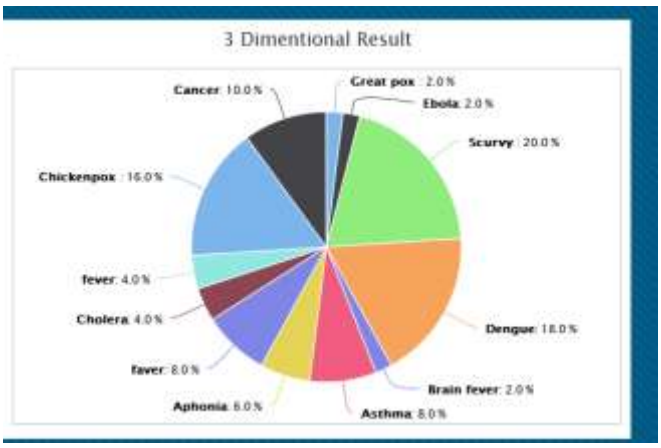

Figure 2. Analysis of dieses

Above graph shows the pictorial view of disease in percentage. This graph of disease can be used for the secondary purpose because it will only display the diseases in percentile ratio not the personal information. Hence the data is anonymized through the differential privacy preserving algorithm and also display such information for the secondary purpose.

VII. CONCLUSION

Data anonymization is one of the superb technologies used for hiding the most sensitive data from the attacker. Differential privacy preserving algorithm is one of suitable mechanism for the statistical data bound in mathematical way. It is very difficult to do the re-identification. Under the DP mechanism convex optimization is used for the purpose of optimization. Every time the user require a key for display the record of patient and it will provide the more security of the patient data and there are no chance to leak the information. Differential privacy mechanism provides the mathematical support for making the information gain .It controls the privacy degradation.

VIII. ACKNOWLEDGMENT

It is a great pleasure to acknowledge those who extended their support, and contributed time and psychic energy for the completion of this project work. At the outset, I would like to thank my project guide Prof. M.M.Naoghare, who served as sounding board for both contents and programming work. Her valuable and skillful guidance, assessment and suggestions

from time to time improved the quality of work in all respects. I would like to take this opportunity to express my deep sense of gratitude towards her, for her invaluable contribution in this project work. I am also thankful to Prof. S.M.Rokade, Head of Computer Engineering Department for his timely guidance, inspiration and administrative support without which my work would not have been in process. I am also thankful to the all staff members of Computer Engineering Department and Librarian, SVIT Chincholi, Nasik. Also I would like to thank my colleagues and friends who helped me directly and indirectly in this Project work. Lastly my special thanks to my family members for their support and co-operation during this Project work.

REFERENCES

[1] Mohamed R. Fouad, Khaled Elbassioni, "A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization", IEEE transaction on Knowledge and Data Engineering July 2014.

[2] M. R. Fouad, K. Elbassioni, and E. Bertino, "Towards a differentially private data anonymization," Purdue Univ., West Lafayette, IN, USA, Tech. Rep. CERIAS 2012-1, 2012.

[3] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially prívate data release for data mining," in Proc. 17th ACM SIGKDD,New York,NY, USA, 2011, pp. 493–501.

[4] M. R. Fouad, G. Lebanon, and E. Bertino, "ARUBA: A risk-utility based algorithm for data disclosure," in Proc. VLDB Workshop SDM, Auckland, New Zealand, 2008, pp. 32–49.

[5] K. M. Elbassioni, "Algorithms for dualization over products of partially ordered sets," SIAM J. Discrete Math., vol. 23, no. 1,pp. 487–510, 2009.

[6] C. Dwork, "Differential privacy: A survey of results," in Proc. Int.Conf. TAMC, Xi˘ San, China, 2008, pp. 1–19.

[7] G. Lebanon, M. Scannapieco, M. R. Fouad, and E. Bertino,"Beyond kanonymity:A decision theoretic framework for assessing Privacy risk," in Privacy in Statistical Databases. Springer LNCS 4302:217U 232, 2006.

[8] C. Dwork, "Differential privacy," in Proc. ICALP, Venice, Italy,2006, pp.1–12.

[9] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality,"in Proc. Int. Conf. VLDB, Trondheim, Norway, 2005,pp. 901–909.

[10] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor,"Our data, ourselves: Privacy via distributed noise generation,"in Proc. 25th EUROCRYPT, Berlin, Germany, 2006, pp. 486–503,LNCS 4004.

[11] A. Frieze, R. Kannan, and N. Polson, "Sampling from log-concave distributions," Ann. Appl. Probab., vol. 4, no. 3, pp. 812–837, 1994.

[12] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in Proc. IEEE ICDE, Washington, DC, USA, 2005, pp. 205–216.

[13] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in Proc. Int. Conf.VLDB, Vienna, Austria, 2007, pp. 758–769.

[14] G. A. Gratzer, General Lattice Theory, 2nd ed. Basel, Switzerland:Birkh¨auser, 2003.

[15] M. Grotschel, L. Lovasz, and A. Schrijver, "Geometric algorithms and combinatorial optimization," in Algorithms and Combinatorics, vol. 2, 2nd ed. Berlin, Germany: Springer, 1993.

[16] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness," J. VLDB, vol. 20, no. 1, pp. 59–81, 2011.

[17] C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in Proc. IEEE ICDE, Washington, DC, USA, 2005, pp. 205–216.