Elysium Journal
Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

A ROBUST DATA OBFUSCATION APPROACH FOR PRIVACY PRESERVING DATAMINING

S.Deebika[1] A.Sathyapriya[2]
[1]PG Student
[2]Assistant Professor
Department of Computer Science and Engineering,
Vivekananda College of engineering for women, Namakkal, India.
[1]Email:deebibtechme@gmail.com
[2]Email:sathysat@gmail.com

Abstract: Data mining play an important role in the storing and retrieving of huge data from database. Every user wants to efficiently retrieve some of the encrypted files containing specific keywords, keeping the keywords themselves secret and not jeopardizing the security of the remotely stored files. For well-defined security requirements and the global distribution of the attributes needs the privacy preserving data mining (PPDM). Privacy-preserving data mining is used to uphold sensitive information from unendorsed disclosure. Privacy preserving data is to develop methods without increasing the risk of misuse of the data. Anonymization techniques: K- Anonymity, L-Diversity, T-Closeness, P-Sensitive and M-invariance offers more privacy options rather to other privacy preservation techniques (Randomization, Encryption, and Sanitization). All these Anonymization techniques only offer resistance against prominent attacks like homogeneity and background. None of them is able to provide a protection against all known possible attacks and calculate overall proportion of the data by comparing the sensitive data. We will try to evaluate a new technique called (n,t)-Closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table.

Index Terms— Anonymization, L-Diversity, PPDM, P-Sensitive, T-Closeness, (n,t)-Closeness.

I. INTRODUCTION

Rapid growth of internet technology have made possible to make use of remote communication in every aspects of life. As well as the increase of technology, privacy and security is needed in electronic communications became warm issues. Security to sensitive data against unofficial access has been a long term goal for the database security study group of people. Data mining consists of number of techniques for manufacture automatically and entertainingly to retrieve the information from the large amount of database which consists of sensitive information too. Privacy is vital issue in transferring of sensitive information from one spot to another spot through internet.

Most considerably, in hospital, in government administrative center and in industries; there is need to establish privacy for sensitive information or data to analyze and future processing on it from other departments. Various organizations (e.g., Hospital authorities, industries and government organizations etc) releasing person thorough data, which called as micro data. They provide information of privacy of individuals. Main aspire is to protect information simultaneously to produce external knowledge.

The table consist of micro data is called Micro table [6]. i) identifiers-Uniquely identified attributes are called as identifiers. e.g., Social Security number. ii) Quasi-identifiers -adversary of attribute may already known and taken together can potentially identify an individual e.g., Birth date, Sex and Zip code. iii) Sensitive attributes - adversary of attribute is unknown and sensitive. e.g., Disease and Salary. [3] are the three tupules.

Sensitive information is fragment different from secret and confidential. Secret information means Passwords, pin codes, credit card details etc. The sensitive information mostly linked to diseases like HIV, Cancer, and Heart Problem etc.

II. RELATED WORKS

The main aim of the privacy preserving is to create method and techniques for the prevention of misusage of sensitive data. The techniques are proposed for altering the original data to carry out privacy. The alteration may not affect the original data and to improve the privacy on it. Various methods of privacy can prevent unauthorized usage of sensitive attribute. Some of the Privacy methods [11][4] are Anonymization, Randomization, Encryption, and Data Sanitization. Extending of this many advanced techniques are proposed, such as p-sensitive k-anonymity, (α, k)-anonymity, l-diversity, t-closeness, M-invariance, Personalized anonymity, and so on. For multiple sensitive attribute[7], there are three kinds of information disclosure.

i) Identity Disclosure: An individual is linked to a particular record in the published data.
ii) Attribute Disclosure: When sensitive information regarding individual is disclosed known as Attribute Disclosure.

iii) Membership Disclosure: When information regarding individual's information is present in data set and it is not disclosed.

When the micro data is published the various attacks are occurred like record linkage model attack and attribute linkage model attack. To avoid these attacks the different anonymization techniques was introduced.

We did many surveys on anonymization [8] techniques. They are explained below.

A. K-Anonymity

K-anonymity is a property possessed by certain anonymized data. The theory of *k*-anonymity was first formulated by L. Sweeney[12] in a paper published in 2002 as an attempt to solve the problem: "Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re identified while the data remain practically useful."[9][10]. A release of data is said have the *k*-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release.

Methods for k-anonymization

In the framework of *k*-anonymization problems, a database is a table with *n* rows and *m* columns. Each row of the table represents a record relating to a specific member of a population and the entries in the various rows need not be unique. The values in a mixture of columns are the values of attributes associated with the members of the population. The following table 1 is a non anonymized database consisting of the patient records.

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

S.No	Zip code	Age	Disease
1	4369	29	TB
2	4389	24	Viral infection
3	4598	28	No illness
4	4599	27	Viral infection
5	4478	23	Heart-related

Table 1 Non Anonymized database

The above table 1 has 4 attributes and 5 records in this data. There are two common methods for achieving *k*-anonymity [13] for some value of *k*.

Suppression: In this Suppression method, certain values of the attributes of column are replaced by an asterisk '*'. In the anonymized below table, have replaced all the values in the 'Name' attribute and the 'Religion' attribute have been replaced by a '*'.

Generalisation: In this method, individual values of attributes are replaced by with a broader category. For example, the value '23' by '20 < Age ≤ 30', etc. The below table 2 shows the anonymized database.

K-anonymity model was developed to protect released data from linking attack but it causes the information disclosure. The protection of k-anonymity provides is easy and simple to appreciate. K-anonymity does not provide a shelter against attribute disclosure. Table 2 is Anonymized version of the database are shown below.

S.No	Zip code	Age	Disease
1	43**	20 < Age ≤ 30	TB
2	43**	20 < Age ≤ 30	Viral infection
3	45**	20 < Age ≤ 30	No illness
4	45**	20 < Age ≤ 30	Viral infection
5	44**	20 < Age ≤ 30	Heart-related

Table 2 Anonymized database.

.

Attacks on k-anonymity

In the section, we study about the attacks on k-anonymity. There are two types of attacks. They are Homogeneity Attack and background attack. Table 3 shows two types of attack

Zip	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790**	>=40	Flu
4790**	>=40	Heart Disease
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Homogeneity attack

Bob	
Zip	Age
47678	27

John	
Zip	Age
47673	36

Background Knowledge attack

Table 3 Homogeneity and Background knowledge attack

Homogeneity Attack

Sensitive attributes are lack in *diversity* values. From the above table, we easily conclude that Bob Zip code is

P-ISSN: 2347-4408
E-ISSN: 2347-4734

up to the range of 476** and his age is between 20 to 29.Then finally conclude he is attacked by Heart Disease. It is said to be Homogeneity attack.

Background Knowledge Attack

Attacker has additional background knowledge of other sensitive data.

Restrictions of K-anonymity

- K-anonymity make visible of individuals' sensitive attributes.
- Background knowledge attack is not protected by K-anonymity.
- Plain knowledge of the k-anonymization algorithm can be dishonored by the privacy.
- Applied to high-dimensional data is not possible.
- K- Anonymity cannot protect against Attribute disclosure.

Variants of K-anonymity

A micro data satisfies the p-sensitive k-anonymity [15] property if it satisfies K-anonymity and the number of distinct values for each sensitive attribute is at least p within the same QI. It reduces information loss through anatomy approach.

(α, k) – Anonymity

A view of a table is said to be an (α, k)-anonymization [16] of the table if the view modifies the table such that the view satisfies both k-anonymity and α –deassociation properties with respect to the quasi-identifier.

B.L-diversity

L-diversity is proposed to overcome the short comes of K-anonymity. It is the extension of K-anonymity. L-diversity [1] is proposed by Ashwin Machanavajjhala in the year 2005.An equivalence class has l-diversity if there is l or more well-represented values for the sensitive attribute. A table is said to be l -diverse if each equivalence class of the table is l-diverse. This can guard against by requiring "many" sensitive values are "well-represented" in a *q** block (a generalization block).

Attacks on l-diversity

In this section, we study about two attacks on l-diversity [2]: the Skewness attack and the Similarity attack.

Skewness Attack

There are two sensitive values, they are HIV positive (1%) and HIV negative (99%).Serious privacy risk Consider an equivalence class that contains an equal number of positive records and negative records *l*-diversity does not differentiate Equivalence class.

Equivalence class 1: 49 positive + 1 negative;
Equivalence class 2: 1 positive + 49 negative.

L-diversity does not consider the overall distribution of sensitive values.

Similarity Attack

When the sensitive attribute values are distinct but also semantically parallel, an adversary can learn important information. Table 4 shows similarity attack.

Zip code	Age	Salary	Disease
476**	2*	20k	Gastric ulcer
476**	2*	30k	Gastric
476**	2*	40k	Stomach cancer
479**	>=4	100k	Gastric
476**	>=4	60k	Flu
476**	3*	70k	Bronchitis

Similarity attack

Bob	
zip	Age
47678	27

Table 4. Similarity attack.

As conclude from table, Bob's salary is in [20k, 40k], which is relative low. Bob has some stomach-related disease.

Variant of L-diversity

Distinct *l*-diversity

Each equivalence class has at least *l* well-represented sensitive values. It doesn't prevent the probabilistic inference attacks. e.g., In one equivalent class, there are ten tupules. In the "Disease" area, one of them is "Cancer", one is "Lung Disease" and the remaining eight are "Kidney failure". This satisfies 3-diversity, but the attacker can still affirm that the target person's disease is "Kidney failure" with the accuracy of 80%.

Entropy l-diversity

Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough.
The entropy of the entire table may be very low. This leads to the less conservative notion of l- diversity.

Recursive *(c,l)*-diversity

The most frequent value does not appear too frequently.

Restrictions of L-diversity

- It prevents Homogeneity attack but l-diversity is insufficient to prevent attribute disclosure.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

- L-diversity is unnecessary and difficult to achieve for some cases.

- A single sensitive attribute two values: HIV positive (1%) and HIV negative (99%) very different degrees of sensitivity.

C. T-closeness

The t-closeness [14] model was introduced to overcome attacks which were possible on l-diversity (like similarity attack). L-diversity model uses all values of a given attribute in a similar way (as distinct) even if they are semantically related. Also not all values of an attribute are equally sensitive. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. It requires that the earth mover's distance between the distribution of a sensitive attribute within each equivalence class does not differ from the overall earth movers distance of the sensitive attribute in the whole table by more than a predefined parameter t.

Restrictions of t-closeness

T-closeness is an effective way when it is combined with generalizations and suppressions or slicing[5]. It can lost co-relation between different attributes because each attribute is generalized separately and so we lose their dependencies on each other. There is no computational procedure to enforce t-closeness. If we consider very small utility of data is damaged.

III. PROPOSED WORK

(n,t) –CLOSENESS

The (n, t)-closeness principle: An equivalence class E_1 is said to have (n, t)-closeness if there exists a set E_2 of records that is a natural superset of E1 such that E2 contains at least n records, and the distance between the two distributions of the sensitive attribute in E_1 and E_2 is no more than a threshold t. A table is said to have (n, t)-closeness if all equivalence classes have (n, t)-closeness. (n,t) -Closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table.

S.No	Zip Code	Age	Disease	Count
1	47696	29	pnemonia	100
2	47647	21	Flu	100
3	47602	28	Pnemonia	200
4	47606	23		200
5	47952	49	Pnemonia	100
6	47909	48	Flu	900
7	47906	47	Pnemonia	100
8	47907	45	Flu	900
9	47603	33	Pnemonia	100
10	47601	30	Flu	100
11	47608	35	Pnemonia	100
12	47606	36	Flu	100

Table 5 Original patient data

In the above definition of the (*n*, *t*)-closeness principle, the parameter *n* defines the breadth of the observer's background knowledge. Smaller *n* means that the observer knows the sensitive information about a smaller group of records. The parameter *t* bounds the amount of sensitive information that the observer can get from the released table. A smaller t implies a stronger privacy requirement

S.No	ZIP Code	Age	Disease	Count
1	476**	2*	Pnemonia	300
2	476**	2*	Flu	300
3	479**	4*	Pnemonia	100
4	479**	4*	Flu	900
5	476**	3*	Pnemonia	100
6	476**	3*	Flu	100

Table 6 An Anonymous Version of table 5

The intuition is that to learn information about a population of a large-enough size (at least *n*). One key term in the above definition is "natural superset". Assume that we want to achieve (1000, 0.1)-closeness for the above example. The first equivalence class E_1 is defined by (zip code="476**", $20 \leq$ Age ≤ 29) and contains 600 tuples. One equivalence class that naturally

.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

contains it would be the one defined by (zip code= "476**", 20 ≤ Age ≤ 39). Another such equivalence class would be the one defined by (zip code= "47***", 20 ≤ Age ≤29). If both of the two large equivalence classes contain at least 1,000 records, and E_1's distribution is close to (i.e., the distance is at most 0.1) either of the two large equivalence classes, then E_1 satisfies (1,000, 0.1)-closeness. In fact, Table 6 satisfies (1,000, 0.1)-closeness. The second equivalence class satisfies (1,000, 0.1)-closeness because it contains $2,000 > 1,000$ individuals, and thus, meets the privacy requirement (by setting the large group to be itself).

The first and the third equivalence classes also satisfy (1,000, 0.1)-closeness because both have the same distribution (the distribution is (0.5, 0.5)) as the large group which is the union of these two equivalence classes and the large group contains 1,000 individuals. Choosing the parameters *n* and *t* would affect the level of privacy and utility. The larger *n* is and the smaller *t* is, one achieves more privacy and less utility.

IV. EXPERIMENTAL SETUP

We did a sample experiment to check the efficiency of the new privacy measure. Here, a sample graph is shown in fig 1.We compared our different techniques with the proposed model and gets the sample graph with efficient manner. We use parameter number of datasets and privacy degree. In this, datasets are given as sample input and getting privacy with the efficient manner as an output.

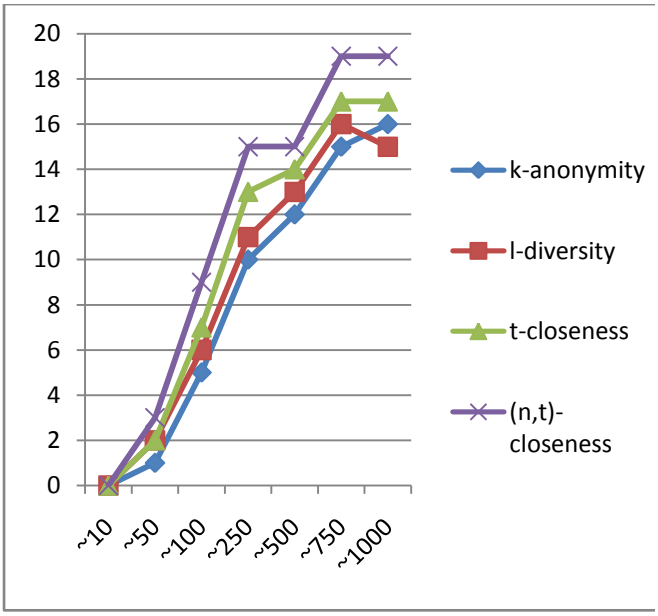

Fig 1 Comparison of different anonymization technique with number of datasets and privacy efficiency

V. CONCLUSION

This paper presents a new approach called (n,t)-Closeness to privacy-preserving micro data publishing. We explained detail about the related works and the drawbacks of anonymization techniques. The new novel privacy technique has overcome the drawbacks of Anonymization technique and generalization and suppression too. It provides security and proportional calculation of data. We illustrate how to calculate overall proportion of data and to prevent attribute disclosure and membership disclosure. We have explained and compared between different types of Anonymization. Our experiments show that (n,t)-Closeness preserves better data utility than Anonymization techniques .

VI. REFERENCES

[1] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. Available at http://www.cs.cornell.edu/_mvnak, 2005.

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam, ℓ-Diversity: Privacy Beyond k-Anonymity 2006.

[3] Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitris Papadias.K-Anonymity in the presence of External database, IEEE Transactions on Knowledge and Data Engineering, vol.22, No.3, March 2010.

[4] Gayatri Nayak, Swagatika Devi, "A Survey on Privacy Preserving Data Mining: Approaches and Techniques", India, 2011.

[5] Li, N. Li, T. Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007: 106-115.

[6] In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2006), pages 754 – 759.

[7] Inan.A,Kantarcioglu.M,and Bertino.e, "Using Anonymized Data for Classification," Proc. IEEE 25th Int Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[8] Li T. and Li N. (2007), Towards Optimal k-Anonymization, Elsevier Publisher, CERIAS and Department of Computer Science, Purdue University, 305 N. University Street, West Lafayette, IN 47907-2107, USA.

[9] L. Sweeney. "Database Security: k-anonymity". Retrieved 19 January 2014.

[10] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

[11] R. Agrawal, R. Srikant, "Privacy-Preserving Data Mining", ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.

[12] Sweeney.L, k-anonymity: a model for protecting privacy. International Journal on Uncertainty,Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[13] Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression.International Journal of Uncertainty, Fuzziness and Knowledge-Based System, 10(5) pp. 571-588, 2002.

[14] t-Closeness: Privacy Beyond k-Anonymity and l – Diversity ICDE Conference, 2007, Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian.

[15] Truta, T.M. and Bindu, V. (2006) Privacy Protection: P-Sensitive K-Anonymity Property. In Proceedings of the Workshop on Privacy Data Management, bwith ICDE 2006, pages 94.

[16] Wong, R.C.W., Li, J., Fu, A.W.C., and Wang, K. (2006) (α, k)-Anonymity:An Enhanced k-Anonymity Model for PrivacyPreserving Data Publishing.