iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** III **Month of publication:** March 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Preserving Privacy of Data Using K-Anonymisation and T-Closeness

Anu Rinny Sunny
PG Scholar, *Department of Computer Science and engineering,*
Avinashilingam Institute for Home Science and Higher Education for Women - Coimbatore

Abstract – In recent years, the growth of Electronic Health Record (EHR) technology has increased the amount of clinical data being electronically available. Medical data, scientific documents, and also digitalised patient health records, are important sources for clinical research. The cause for the leakage of private data seems to be the traditional data mining techniques and algorithms which operates on the original data set itself. Privacy preservation for individual's medical information is vital before taking it to the secondary stages of research. To overcome these challenges imposed by traditional data mining techniques, privacy-preserving data mining (PPDM) has become one of the newest trends in privacy and security in data mining research. Modification of the data needs algorithms to be developed which, however should not compromise the privacy of the original data. This stands as the main aim of privacy preservation. Different techniques such as K anonymity, L-Diversity and T-closeness are used to preserve privacy of sensitive data.
Keywords: PPDM-Privacy Preserving Data Mining, K-anonymity, L-Diversity, T-closeness, anonymisation

I. INTRODUCTION

Data mining aims to mine knowledge from large databases to get useful information that is embedded in data. Private and sensitive information about users needs to be extracted which forms the basis for many data mining applications. The disclosure of patient electronic health records imposes severe threat to the people. This brings trust issues from patients' side which makes it further difficult for the data to be included for research. Data mining algorithms run on confidential data should be kept secret even to the party running the algorithm. Privacy preserving data mining is divided into two they are i) sensitive information such as names, addresses, age and so on, should be modified from the original database, so that the privacy of the data recipient is not compromised, ii) sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because those data can indirectly lead to privacy attacks. So, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. The users' personal information is referred to individual privacy preservation and the latter is referred to collective privacy preservation. To avoid these problems, privacy preserving methods providing additional privacy guarantees over the basic data de-identification must be developed. Various privacy preserving techniques like disassociation-based approach that enforces K-anonymity, L-diversification, and T-closeness with low information loss are implemented. Publishing transaction data safely requires the elimination of two types of potential privacy leak, namely identity and sensitive information disclosure. Identity disclosure occurs when an individual is linked to their transaction in the published data. Unfortunately, simply de-identifying transactions (i.e., removing personal identifiers) is not enough to prevent this type of disclosure. Disclosure of personal information, on the other hand, occurs when an individual is linked not to a transaction, but to a set of items that is considered to be sensitive. Different techniques are discussed in order to preserve privacy in medical datasets. These are three categories of attributes in micro data. In both the anonymization techniques first identifiers are removed from the data and then partitions the tuples into buckets. Generalization transforms the quasi-identifying values in each bucket into less specific and semantically constant so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SA values from the QI values by randomly permuting the SA values in the bucket .The anonymized data consist of a set of buckets with permuted sensitive attribute values. The identity of patients must be protected when patient data is shared.

©IJRASET 2015: All Rights are Reserved

Volume 3 Issue III, March 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

II. K-ANONYMITY

K-anonymity is a classic model, which generalizes and suppresses portions of the released data so as to prevent joining attacks. By this unique identification of an individual from a group of size k is prevented. In the k-anonymous tables, a data set is said to be k-anonymous ($k \geq 1$) if each record in the data set is indistinguishable from at least (k-1) other records within the same data set. If the value of k is larger, privacy is highly protected. K-anonymity can ensure that individuals cannot be uniquely identified by linking attacks. Identity disclosure may occur even when data, devoid of information that explicitly identifies individuals, are published. This is because released data may still be linked to external data containing individuals' identities, through seemingly innocuous attributes, such as age or gender. To prevent identity disclosure in such cases, k-anonymity is proposed. A relational table satisfying k-anonymity ensures that the probability of linking an individual to their record, based on potentially linkable attributes, is upper bounded by 1/k. To perform identity disclosure, an attacker must possess three types of knowledge (i) a patient's identity,(ii) a set of diagnosis codes, and (iii) whether a patient is included in the published dataset. Knowledge of the first two types can come in the form of background knowledge or may be solicited by exploiting external data sources. At the same time, knowledge of the third type is obtainable through interaction with data subjects. K-anonymity safeguards the data from identity disclosure, but it does not protect the data from background knowledge attack. K-anonymity prevents the disclosure of the patient details that are caused due to linking attacks. Further improved model of k-anonymity is the k^m-anonymization model. With this model the probability of identifying the patient reduces, as it becomes less vulnerable to the attacks from the external sources. The published dataset must have no information loss and should possess high utility gain for which utility factors are introduced for K-anonymity. For the research purpose 2 datasets are taken into consideration Adult dataset and Medical dataset. The adult dataset is taken from UCI repository which has 815 instances. K anonymity is performed on 5 quasi identifiers and 2 sensitive attributes. The sensitive attributes are relationship and occupation. The quasi identifiers are taken to be education number, education, work class, hours/week and marital status. The medical dataset contains comparatively lesser instances than the adult dataset. The medical data set is collected from true patient details in order to work with original datasets to understand the utility differences. The K-anonymity is done with bucket size to be 6. The medical data set is collected from true patient details in order to work with original datsets to understand the utility differences. The K-anonymity is done with bucket size to be 6. K-anonymity prevents linkage attacks but does not prevent attribute disclosure. In order to achieve higher levels of privacy different techniques are used such as ℓ-diversity and T-closeness.

AGE	WORKC...	FNLWGT	EDUCA...	EDUCA...	MARRIT...	OCCUP...	RELATI...	RACE	SEX	CAPITA...	CAPITAL...	HOURS...	NATIVE	INCOME
26	Federa...	352768	HS-grad	9	Divorced	Level 5	Own-c...	White	Female	0	0	40	United-...	<=50K
19	Private	356717	Some-...	10	Never-...	Level 5	Own-c...	White	Female	0	0	25	United-...	<=50K
20	Private	138352	HS-grad	9	Never-...	Level 3	Other-r...	White	Male	0	0	30	United-...	<=50K
33	Federa...	94193	HS-grad	9	Divorced	Level 5	Unmar...	White	Female	0	0	40	United-...	<=50K
43	State-g...	206927	HS-grad	9	Marrie...	Level 1	Husba...	White	Male	0	0	40	United-...	<=50K
19	Private	517036	HS-grad	9	Divorced	Level 4	Not-in-f...	White	Female	0	0	40	El-Salv...	<=50K
37	Local-...	31023	Some-...	10	Marrie...	Level 3	Husba...	White	Male	0	0	40	United-...	<=50K
55	Private	174478	10th	6	Never-...	Level 3	Not-in-f...	White	Male	0	0	29	United-...	<=50K
43	Federa...	32016	Masters	14	Marrie...	Level 1	Husba...	White	Male	0	0	40	United-...	>50K
42	Private	155469	Assoc-...	12	Widow...	Level 2	Unmar...	White	Female	0	0	24	United-...	<=50K
18	Private	126125	HS-grad	9	Never-...	Level 3	Own-c...	White	Male	0	0	20	United-...	<=50K
31	Private	167725	Bachel...	13	Marrie...	Level 1	Husba...	Asian-...	Male	15024	0	48	Philipp...	>50K
35	Private	214891	HS-grad	9	Marrie...	Level 3	Husba...	Other	Male	0	0	40	Domini...	<=50K
31	Private	227325	Some-...	10	Never-...	Level 4	Not-in-f...	White	Male	0	0	40	United-...	<=50K
26	Private	213081	HS-grad	9	Never-...	Level 5	Unmar...	Black	Female	0	0	40	Jamaica	<=50K
32	Private	107218	Bachel...	13	Never-...	Level 5	Not-in-f...	Asian-...	Male	0	0	40	United-...	<=50K

Fig.1 (a) Original Dataset

©IJRASET 2015: All Rights are Reserved

III. ℓ-DIVERSITY

The technique ℓ-diversity is proposed in order to handle the shortcomings of the k-anonymity, where the technique does not ensure protection against attribute disclosure. l-diversity solves the problem of attribute disclosure by replacing the dataset by ℓ-diversified values for each sensitive attribute. The advantage of l-diversity over k-anonymity is that l-diversity provides higher level of protection to external attacks more than what k-anonymity provides, since k-anonymity does not offer protection against linking attacks .In this model, an equivalence class is said to have l-diversity if there are at least l well-defined values for the sensitive attribute. While publishing data the sensitive information that can infer to the patient indirectly should be reduced in order to prevent the linking and combination attacks. The attacker may be able to mine the relationship between different attributes an find the patient when it comes to larger datasets where there are numerous attributes pertaining to the identity of the patient. Several privacy principles were proposed to prevent this threat. l-diversity requires each group of records to have at least l "well-represented" sensitive values. The dataset contains 152 data items for which bucket size is taken to be ℓ value which i have taken it as 6. The bucket must show ℓ diversified values. This is done for the sensitive attributes. Quasi identifiers are not ℓ-diversified.

IV. T-CLOSENESS

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 value t. A table is said to have t-closeness if all equivalence classes have t-closeness. Because there are semantic relationships among the attribute values, and different values have very different levels of sensitivity. The t-closeness of a table should be high in order to get the highest levels of protection form security threats such as identity disclosure and background knowledge attacks. Earth movers' distance formula is used to find the distance between the attributes. A group of clusters are formed to apply the formula and the distance is calculated. After the ℓ-diversification there may be indirect knowledge attacks with quasi identifiers. The privacy factors and utility factors are used for the t-closeness. The earth mover's formula prevents similarity attack and provides higher levels of privacy protection to the dataset.

AGE	WORKC...	FNLWGT	EDUCA...	EDUCA...	MARRIT...	OCCUP...	RELATI...	RACE	SEX	CAPITA...	CAPITAL...	HOURS...	NATIVE	INCOME
19-33	Federal...	352768	HS-grad	9-9	Divorced	Level 4	Own-ch...	White	Female	0	0	25-40	United-...	<=50K
19-33	Private	356717	HS-grad	9-9	Divorced	Level 3	Other-r...	White	Female	0	0	25-40	United-...	<=50K
19-33	Private	138352	HS-grad	9-9	Divorced	Level 1	Not-in-f...	White	Male	0	0	25-40	United-...	<=50K
19-33	Federal...	94193	HS-grad	9-9	Divorced	Level 5	Husband	White	Female	0	0	25-40	United-...	<=50K
19-33	State-gov	206927	HS-grad	9-9	Divorced	Level 1	Unmarr...	White	Male	0	0	25-40	United-...	<=50K
19-33	Private	517036	HS-grad	9-9	Divorced	Level 1	Not-in-f...	White	Female	0	0	25-40	El-Salv...	<=50K
19-35	Private	93518	Some-c...	7-10	Never-...	Level 4	Own-ch...	White	Female	0	0	20-40	United-...	<=50K
19-35	Private	296738	Some-c...	7-10	Never-...	Level 5	Not-in-f...	White	Female	6849	0	20-40	United-...	<=50K
19-35	Private	409230	Some-c...	7-10	Never-...	Level 2	Husband	White	Male	0	0	20-40	United-...	<=50K
19-35	Private	82623	Some-c...	7-10	Never-...	Level 3	Unmarr...	White	Male	0	0	20-40	United-...	<=50K
19-35	Self-em...	317660	Some-c...	7-10	Never-...	Level 2	Husband	White	Male	7688	0	20-40	United-...	>50K
19-35	Private	148998	Some-c...	7-10	Never-...	Level 2	Husband	White	Female	0	0	20-40	United-...	<=50K
35-54	Private	184378	Bachel...	4-14	Never-...	Level 6	Unmarr...	White	Male	0	0	40-40	United-...	<=50K
35-54	Private	117872	Bachel...	4-14	Never-...	Level 2	Not-in-f...	Black	Male	0	0	40-40	United-...	>50K
35-54	Private	62793	Bachel...	4-14	Never-...	Level 4	Husband	White	Female	0	0	40-40	United-...	<=50K
35-54	Private	161637	Bachel...	4-14	Never-...	Level 1	Other-r...	Asian-...	Male	0	1902	40-40	Taiwan	>50K

Fig.1 (b) T-closeness for multiple sensitive attributes

V. CONCLUSION AND FUTURE WORK

K-anonymity is used to prevent identity disclosure but does not protect the data from linking attacks which lead to the use of another technique called ℓ-diversity. This technique solves the problem of attribute disclosure and preventing the data from being attacked. ℓ

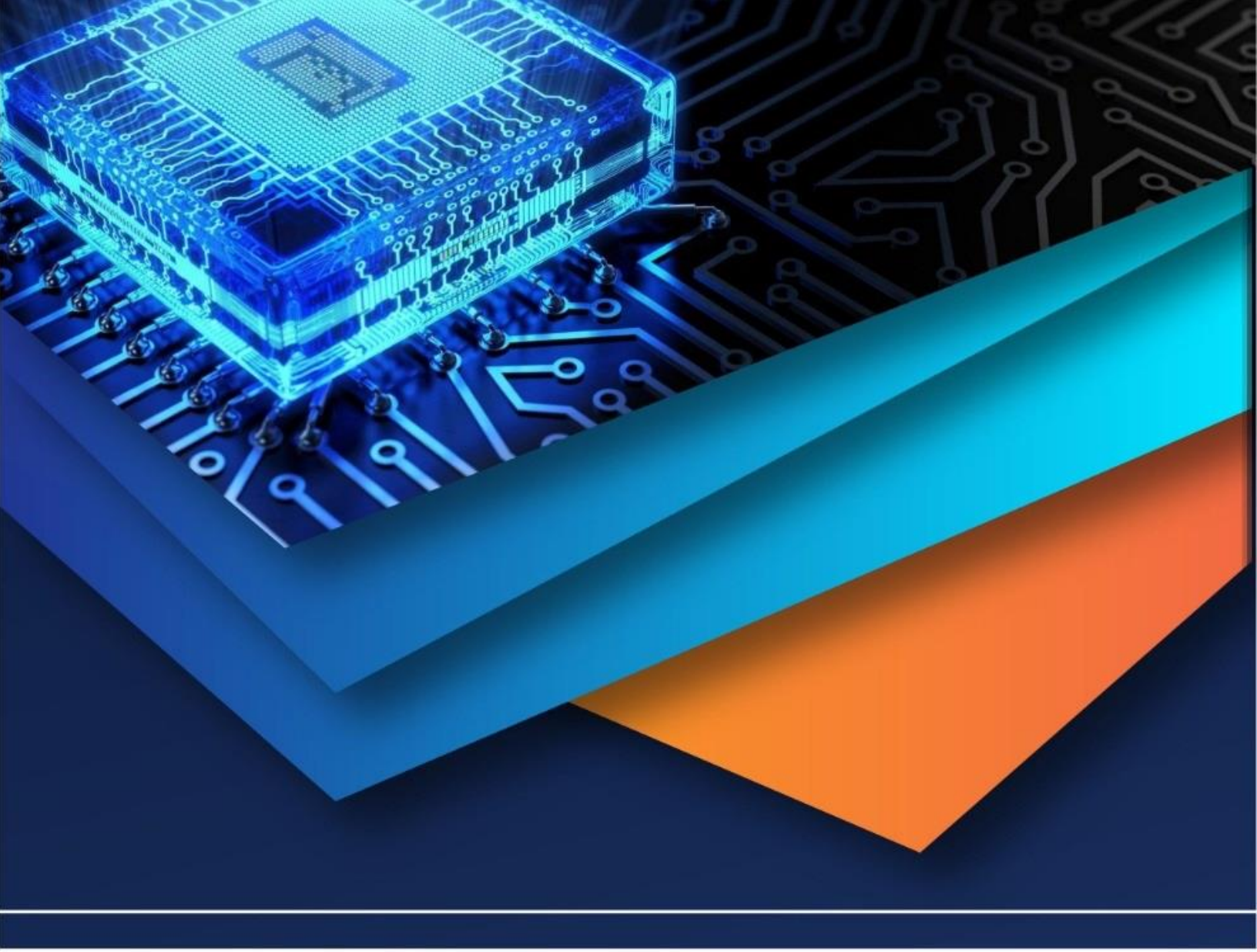
©IJRASET 2015: All Rights are Reserved

-diversity attempts to solve this problem by representing each class by making the class have well-represented values for each sensitive attribute. We have shown that ℓ -diversity has a number of limitations and has proposed another privacy model called t-closeness. Among two distributions the i^{th} values from both are compared and the closeness between them is calculated, keeping the threshold constant. In cases where the threshold values tend to change the next i+1 value among the two distributions are compared and replaced for the same.

REFERENCES

[1]ArisGkoulalas-Divanis, GrigoriosLoukides ,Jimeng Sun "Publishing data from electronic health records while preserving privacy:A survey of algorithms" Journal of Biomedical Informatics 50- 4–19 (2012)
[2]GrigoriosLoukides, ArisGkoulalas-Divanis, "Utility-preserving transaction data anonymization with low information loss" Expert Systems with Applications 39- 9764–9777 (2014)
[3]Javier Parra-Arnau , David Rebollo-Monedero, Jordi Forné, "Measuring the privacy of user profiles in personalized information systems"Future Generation Computer Systems 33 53–63 (2014)
[4]Li Xiong, VaidySunderam, Liyue Fan, SlawomirGoryczka, Layla Pournajaf, "PREDICT: Privacy and Security Enhancing Dynamic Information Collection and Monitoring" Procedia Computer Science 18- 1979 – 1988 (2013)
[5]ManolisTerrovitis and John Liagouris , "Privacy Preservation by Disassociation" Proceedings of the VLDB Endowment, Vol. 5, No. 10 (2012)
[6]Murat Kantarcioglu, Ali Inan , Wei Jiang , Bradley Malin , "Formal anonymity models for efficient privacy-preserving joins" Data & Knowledge Engineering 68 1206–1223 (2009)
[7]Rashid HussainKhokhar , Rui Chen , Benjamin C.M. Fung , Siu Man Lui, "Quantifying the costs and benefits of privacy-preserving health data publishing" Journal of Biomedical Informatics 50 -107–121 (2014)
[8]SoohyungKim , Min Kyoung Sung , Yon Dohn Chung , "A framework to preserve the privacy of electronic health data streams" Journal of Biomedical Informatics 50-95–106 (2014)
[9]Srinivasa L. Chakravarthya, ValliKumariV.a, SarojiniCh , "A Coalitional Game Theoretic Mechanism for Privacy Preserving Publishing Based on k-Anonymity" Procedia Technology 6- 889 – 896 (2012)
[10]R.VidyaBanu, N. Nagaveni, "Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario" Information Sciences 232 437–448 (2013)
[11]Yeye He and Jeffrey F. Naughton , "Anonymization of SetValued Data via TopDown, Local Generalization" VLDB '09, August 2428, Lyon, France(2009)
[12]Xin Jin a, NanZhang ,Gautam Das , "ASAP: Eliminating algorithm-based disclosurein privacy-preserving data publishing" Information Systems 36- 859–880 (2011)
[13] Y. Rubner, C. Tomasi, and L. J. Guibas. "The earth mover's distance as a metric for image retrieval". Int. J. Comput. Vision, 40(2):99–121, 2000
[14] V. S. Iyengar. "Transforming data to satisfy privacy constraints" .In Proc. 8th ACM KDD, pages 279–288, 2002.
[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam._-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

©IJRASET 2015: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)