A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity

Bhavana Abad (Khivsara)
M.E. CSE
MIT, Aurangabad

Kinariwala S.A.
Asst Prof. in computer dept
MIT,Aurangabad
India

ABSTRACT

K-anonymity is one of the easy and efficient technique to achieve privacy preserving for sensitive data in many data publishing applications. In k-anonymity techniques, all tuples of releasing database are generalized to make it anonymize which lead to reduce the data utility and more information loss of publishing table. This paper firstly proposes a Sensitivity Based Tuple Anonymity Method. In this method first we consider the sensitivity of values in sensitive attribute and then only tuples having sensitive values are generalized, and the other tuples can be directly published. Experiment results on the Adult Database show the proposed methods not only can improve the accuracy of the publishing data, but also can preserve privacy

General Terms

Privacy preserving in data mining.

Keywords

Privacy preserving, k-Anonymity, sensitive tuple

1. INTRODUCTION

Many organizations collect and hold large volumes of data like credit card companies, real estate companies, hospitals and search engines. They would like to publish the data for the purposes of data mining. When these organizations publish data it also contains a lot of sensitive information, so they would like to preserve the privacy of the individuals represented in the data[1]. So manipulate data in order to protect the privacy of the individual, to which micro data release refer, data provider often remove key attributes such as names, addresses. De-identifying data, however, provides no guarantee of anonymity. Released information often contains other data called as Quazi-identifiers such as , birth date, sex, and ZIP code[2], which can be linked to publicly available information to re-identify the individual, thus leaking information that was not intended for disclosure. The large amount of information easily accessible today, makes such linking attacks a serious problem. There can be different ways to achieve the goal of privacy in which the releasing some limited data instead of pre-computed heuristics is a increased flexibility and availability for the users. So in Privacy Preserving Data Mining we look for methods to transform the original data such that heuristics determined from the transformed data are close to original heuristics and the privacy of users is not dying out. One way to achieve this is to have the released information adhere to k-anonymity. Intuitively, k-anonymity states that each release of data must be such that every combination of values of released attributes that are also externally available and therefore vulnerable for linking can be indistinctly matched to at least k respondents. In k-anonymity the attributes of tables are classified in three classes. First is key attribute which is generally the name and it is removed at time of releasing, second class is quazi identifier which are generally linked with publicly available database to re-identify the individual, these class contains attributes such as gender, age ,post code. Third class is sensitive attribute which is used by researcher and generally published directly.

Table I Classification of Attributes for k-anonymity

Key attribute	Quasi identifier			Sensitive attributes
Name	Gender	Age	Zip code	Diagnosis
John Smith	Male	25	423101	depression
	Male	27	423508	HIV

How quazi identifiers can be used to re-identify individual using linking attack is given in below example. The two tables are given, Table II contains Medical data set and Table III contains voter list which is also available publically. By linking Zip code, Age and Sex of medical table with voter list table attacker can identify that Arjun is suffering from cancer and in this way the privacy of individual is violated. This is happened because the combination of quazi identifiers value is unique in medical data set, if release data in such a way that there is no unique combination for quazi identifiers then this type of re-identification is not possible. This can be done using anonymizing tables.

Table II Medical Data set

ID	ZIPCODE	AGE	SEX	DIAGNOSIS
1	423065	29	M	Heart Disease
2	422036	32	F	Flu
3	423245	38	M	Cancer
4	422035	37	F	HIV
5	423012	47	M	Headache
6	423432	53	F	Viral

Table III Voter List

NAME	ZIPCODE	AGE	SEX
Mohit	423234	49	M
Sunil	466987	35	M
Shyam	423223	28	M
Rohini	424435	41	F
Arjun	423012	47	M
Sangita	423446	33	F

To avoid the identification of records in micro data, the traditional approach is to de-identify records by removing the identifying fields such as name, address, phone number and social security number.

In order to avoid linking attacks using quasi-identifiers, Sweeney [2] proposed the k-anonymity model, where some of the quasi-identifier fields are suppressed or generalized. A table satisfies k-anonymity if every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes. Such a table is called a k-anonymous table. Hence, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks. Table IV shows a 2-anonymous view corresponding to Table II. The sensitive attributes (Diagnosis Result) is retained without change in this example.

Table IV 2-Anonymous view of Table II

ID	ZIP CODE	AGE	SEX	DIAGNOSTIC
1	423***	>25	M	Heart Disease
2	423***	>25	M	Cancer
3	422***	3*	F	Flu
4	422***	3*	F	HIV
5	423***	>40	*	Headache
6	423***	>40	*	Viral

2. RELATED WORK

In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. Samarati [4] presents an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal k-anonymous table. Model such as l-diversity proposed in 2006 by A. Machanavajjhala [5] solve k-anonymity problem. It tries to put constraints on minimum number of distinct values seen within a equivalence class for any sensitive attribute, S. Venkatasubramanian in 2007 [3] present a model called t-closeness was introduced to overcome attacks possible on l-diversity like similarity attack, R. Wong, J. Li, A. Fu, K. Wang [7] propose an (α, k)-anonymity model to protect both identifications and relationships to sensitive information in data were proposed in the literature in order to deal with the problem of k-anonymity. Bayardo and Agrawal [9] present an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal k- anonymous table. Fung et al. [8] present a top-down approach to make a table satisfied k-anonymous LeFevre et al. [6] describes an algorithm that uses a bottom-up technique . Pei[10] discuss the approaches for multiple constraints and incremental updates in k-anonymity. However the traditional k-anonymity models take consider that the all values of the sensitive attributes are sensitive and need to be protected. In fact, the values which will breach individuals' privacy are in the minority of the whole sensitive attribute dataset. The previous models lead to excessively generalize and more information loss in publishing data.

The work presented in this paper mainly considered the tuples which are really sensitive and need to be preserving the privacy of individual are only generalized and anonymized.

3. CONCEPT AND PROBLEM DEFINATION

The objective of proposed model is as follows

- **Privacy** – To provide the individual data privacy by generalization in such away that data re-identification can not be possible
- **Data utility** - The goal is to eliminate the privacy breach (how much an adversary learn from the published data) and increase utility (accuracy of data mining task) of a released database. This is achieved by generalizing quasi-identifiers of only those tuples having high sensitive attribute values.
- **Minimum information loss** – The loss of information is minimized by giving sensitivity level for sensitive attribute values, and tuples which belongs to high sensitive level are only generalized rest of the tuples are released as it is.

The traditional k-anonymity models take all tuples in publishing table T as sensitive tuples. So they are to be generalized and the publishing data lost a lot of useful information. In this proposed method, firstly an algorithm called Sensitivity Based Tuple Anonymity Method (SBTAM) is to be presented. The kernel idea is to protect individuals' privacy as well as only the high sensitive tuples should be generalized with a satisfied parameter k. The other tuples should not be generalized and can be published directly.

Basic Notation. Let $T\{K_1,K_{2..},K_j\ Q_1,Q_2,..,Q_p,S\}$ be a table. For example, T is a medical dataset. Let $Q_1,\ldots,Q_P$ denote the quasi-identifier specified by the application (administrator). Let S denote the sensitive attribute. A sensitive attribute is an attribute whose value for some particular individual must be kept secret from people who have no direct access to the original data. Let K_j denote the key attributes of T which is to be removed before releasing a table. t[X] denote the value of attribute X for tuple t. |T| denote the number records of T.

Definition 1. (Quasi-identifier): A set of non-sensitive attributes {Q1, …,Qp} of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify (can be called as candidate key) at least one individual in the general population.

Definition 2. (k-Anonymity): A table T satisfies k-anonymity if for every tuple t of T there exist (k-1) other tuples ti1, ti2, tik-1 ∈ T such that t[F] = ti1 [F] =ti2 [F] = … = tik−1 [F] for all F∈ QI.

Definition 3. (Sensitive-values Set): A Set A consists of values which the user selected as most sensitive values from set S. Denote by A.

Definition 4. (Sensitive tuple): Let t ∈ T, if t[S] ∈A, we called t is a sensitive tuple.

3.1 Model and Algorithm

Algorithm for Sensitivity Based Tuple Anonymity Method (SBTAM) :

Input: A table T, quasi-identifier attributes Q,
Sensitive values A, Anonymity parameter k

Output: Releasing Table T*

Step 1: Select database and Table T as shown in Fig 1 below

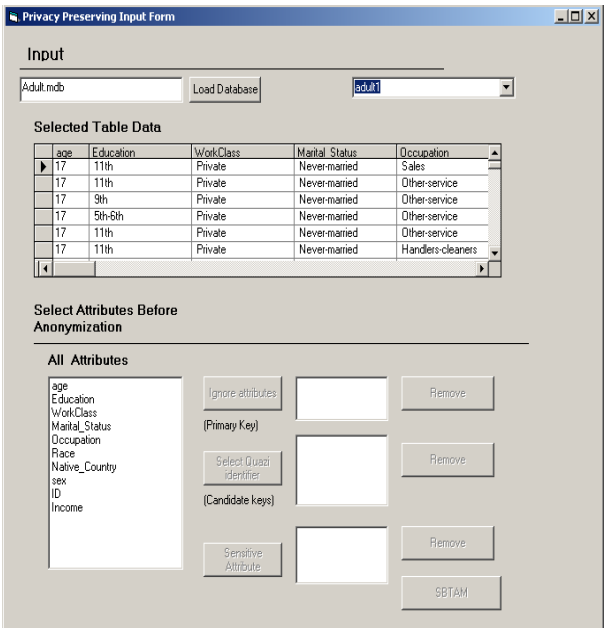

Fig 1 Select Input Table

Step 2: Select Key attribute, Quazi-identifier attribute and Sensitive Attribute from give n attribute list as shown in Fig 2

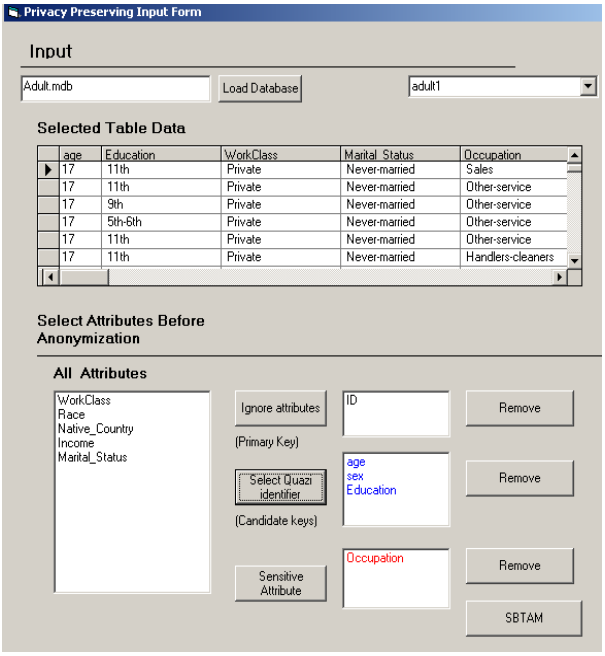

Fig 2 Select Key, Quazi and sensitive Attribute

Step 3: Select the set of most sensitive values A from list of all sensitive values that is to be preserved as shown in Fig 3

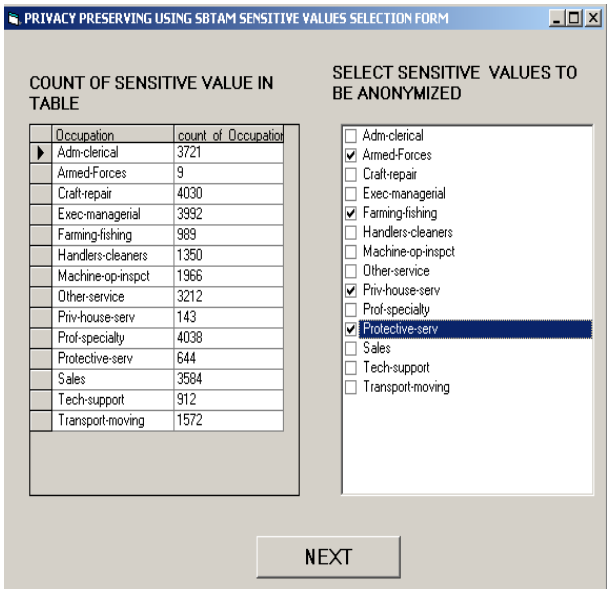

Fig 3 Select set of Sensitive Values – A

Step 4: For each tuple whose sensitive value belongs to set A
If t[S] ∈ A then move all these tuples to Table T1 and rest to table T2.

Step 5: Find the statistics of quazi attributes of table T1 i.e. distinct values for that attribute and total no of rows having that value as shown in fig 4.

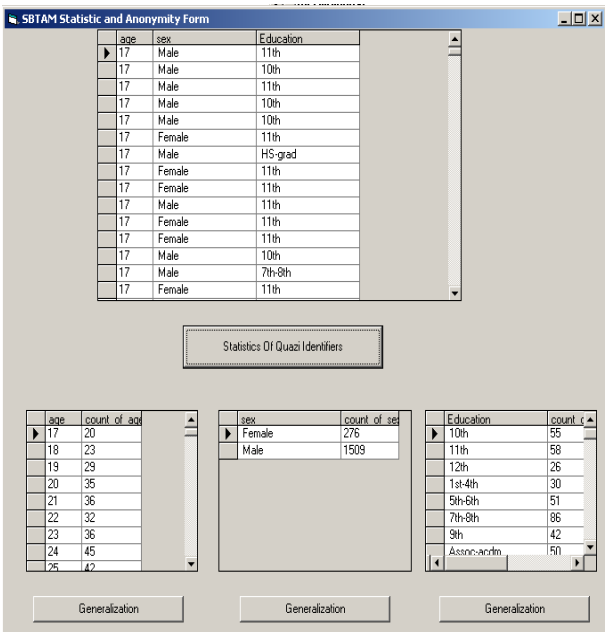

Fig 4 Statistics of Quazi identifiers

Step 6: Apply generalization on quazi identifiers of table T1 to make it k- anonymized

Step 7 : Append rows of table T1 and table T2.
T*=T1+T2 which is table ready to release.

The proposed output of this method is like below shown in Table V. Only tuples containing sensitive values from set A is generalized rest all are published directly so information loss is reduced and so data is utilized more precisely

Table V Sensitivity based tuple anonymization

Zip Code	Age	Sex	Diagnostic Result
423061	27	M	Flu
453063	26	F	Headache
453052	32	F	Skin Infection
493051	36	F	Heart Disease
41305*	4*	*	Cancer
41305*	4*	*	HIV

4. EXPERIMENTAL RESULTS

This method is computed on the Adult Database from the UCI Machine Learning Repository [11]. The Adult Database contains 32561 tuples from US Census data. After preprocessing data and removing tuples containing missing values 30162 tuples are selected. This database containts 11 attributes from that only 5 attributes are used. From that four attributes are consider as quazi-identifier and one attribute 'occupation' as a sensitive attribute. Machine-op-inspct, 'Tech-support', 'Protective-serv' as privacy values which need to be protected.

Table VI provides a brief description of the data including the attributes used in method, the number of distinct values for each attribute, the type of generalization that was used for Quazi identifier attributes and the height of the generalization hierarchy for each attribute.

Table VI Description of Adults Data Set

	Attribute	Distinct	Generalizations	Height
1	Age	74	10-,20-,30	4
2	Marital Status	7	Taxonomy Tree	3
3	Race	5	Taxonomy Tree	2
4	Sex	2	Person	1
5	Occupation	14	Sensitive Attribute	

The precision of publish table R describes as equation (1). Where n=|T|, m=|QI|, $|DGH_{Ai}|$ denote the height of the generalization hierarchy for attribute A_i, h_{ji} denote the height of the generalization hierarchy for tuple tj in attribute A_i.

$$Prec(R) = 1 - \frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\left|\frac{h_{ij}}{DGH_{Ai}}\right|}{n.m} \qquad (1)$$

Fig. 5 shows the precision of publishing data varies along with anonymity parameter k when |QI| is 3.

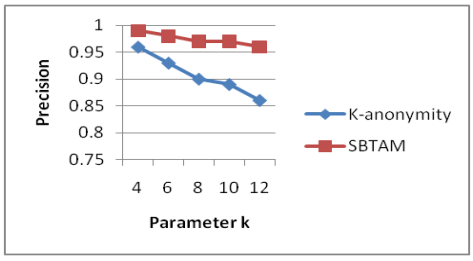

Fig 5 Precision along with K

From fig 5 it is clear that the traditional k-anonymity algorithm yields more information loss because all tuples are involved in generalized. Only sensitive tuples are generalized in SBTAM.

5. CONCLUSION

As concluding remark there are following main issues or threats in traditional k-anonymity privacy preserving algorithms which consider all of sensitive attribute values at same level and apply generalization on all , this leads to some issues like

1. Information Loss
2. Data Utility
3. Privacy measure

So there is a need to develop a method which provide the privacy with minimum information loss and maximum data utility and this paper present a new k-anonymity model based on sensitivity of sensitive attribute in tuples called SBTAM using which information loss is reduced as only sensitive tuples are anonymized, data utility is increased and privacy of individual is also preserved.

6. REFERENCES

[1] Models and Algorithms : Privacy-Preserving Data Mining Springer 2008 : Charu Aggarwal , Philip Yu

[2] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty Fuzziness Knowledge based Systems,10(5), pp 557-570. 2002

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

[4] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):10101027.2001

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam , "l-Diversity: Privacy beyond k-anonymity",In: Proceedings of the IEEE ICDE 2006

[6] K LeFevre, D DeWitt , R Ramakrishnan. Incognito:Efficient full domain k-anonymity Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore , Maryland , 2005: 49-60 .

[7] R. Wong, J. Li, A. Fu, K. Wang. (α, k)-anonymity: an enhanced k-anonymity model For privacy preserving data publishing. KDD 2006:754-759

[8] B. Fung, K. Wang, P. Yu. Top-down specialization for information Conference on Data Engineering (ICDE05),pp:205-216

[10] R. Bayardo and R. Agrawal. Data privacy through optimal k- anonymity. In Proceedings of the 21st International onference on Data Engineering (ICDE), pp:217-228,Tokyo,Japan,2005

[11] J Pei , J Xu , Z. B Wang , W Wang, K Wang. Maintaining k- anonymity against incremental updates//Proceedings of the 19th International Conference on Scientific and Statistical Database

[12] U.C.Irvine Machine Learning Repository, http://www.ics.uci.edu/mlearn/mlrepository.html