Sensitive Attributes based Privacy Preserving in Data Mining using k-anonymity

Nagendra kumar.S
Department of IT
Rajalakshmi Engineering College,
Chennai

Aparna.R
Department of CSE
Prince Dr.K.Vasudevan
College of
Engineering, Chennai

ABSTRACT
Data mining is the process of extracting interesting patterns or knowledge from huge amount of data. In recent years, there has been a tremendous growth in the amount of personal data that can be collected and analyzed by the organizations. Organizations such as credit card companies, real estate companies and hospitals collect and hold large volumes of data for their research purposes. E.g. National Institute of health. When these organizations publish data containing a lot of sensitive information. The importance of sharing data for research and knowledge discovery has been well-recognized. However, sharing data that contains sensitive personal information, such as insurance data, medical record, etc across organization boundaries can raise serious privacy concerns. There is a need to preserve the privacy of the individuals in data set . K-anonymity is one of the easy and efficient techniques to achieve privacy in many data publishing applications. In k-anonymity, all tuples of releasing database are generalized to make it anonymize which lead to data utility reduction and more information loss of publishing table. Sensitive attribute based anonymity method is very useful in preserving the privacy of individuals in organization's publication of data. It reduces information loss to the researchers by providing sensitive levels. This method also avoids Homogeneity attack and Background attacks.

Keywords
Privacy preserving, k-Anonymity, sensitive attributes

1. INTRODUCTION

1.1 Data Mining
Data mining is a well-known technique for automatically and intelligently extracting information or knowledge from a huge amount of data. Data mining, the extraction of hidden predictive information from huge databases, is a new technology with great potential to help companies focus on the most important information in their data warehouses. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining. This evolution began when business data was first stored on computers, sustained with improvements in information access, and the newly generated technologies allow users to find the way through their data in real time. Data mining takes this development beyond retrospective data access and navigation to prospective and proactive information delivery.

1.2 Privacy Preserving Data Mining
In recent years, increasing amount of personal data is stored by individual and private corporations because of advances in storage and collection of hardware technology. This has shown the way to improved privacy concerns about the safety of the original data. In order to make an openly accessible system protected, the privacy of data must be ensured. So privacy preserving data mining (PPDM) has become a significant field of research.

Privacy Preserving Data Mining looks for methods to alter the original data such that heuristics determined from the altered data are close to original heuristics and the privacy of users is not fading out. A number of effective methods for privacy preserving data mining have been proposed. But most of these methods might result in information loss and side effects in some extent such as reduction in data utility, downgrading the efficiency of data mining.

1.3 k - anonymity
One way to achieve PPDM 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. It requires that each record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. Table 1 shows the Two anonymous view(k value as 2) of a data set. k-anonymity protects against identity disclosure, but does not protect attribute disclosure which leads to homogeneity attack.

Table 1. Two Anonymous View Of A Data Set

Id	Zipcode	Age	Gender	Diagnostic
1	423***	>25	M	Flu
2	423***	>25	M	Flu
3	4236**	3*	F	Cancer
4	4236**	3*	F	Cancer
5	428***	>40	*	HIV
6	428***	>40	*	HIV

2. LITERATURE REVIEW
Samarati [10] presents an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal k-anonymous table. Sweeny [9] proposed a model where the k-anonymity protection of the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release.

Model such as l-diversity proposed by A. Machanavajjhala [8] in 2006 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 [7] 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 6] 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. John Miller, Alina Campan and Traian Marius Truta [5] specifies about quasi identifiers generalization boundaries and achieving k-anonymity within the imposed boundaries. Limiting the amount of generalization when masking micro data is indispensable for real life datasets and applications. Batya Kenig and Tamir Tassa [2] introduced a method of mining closed frequent generalized records. Experiments show that the significance of algorithm is not limited to the theory of k-anonymization. This achieves lower information loss than the leading approximation algorithms. However the traditional k-anonymity models consider that all values of the attributes are sensitive and need to be protected. In fact, the values which will breach individual's privacy are in the minority of the whole sensitive attribute dataset. The previous models lead to excessive generalization and suppression that leads to more information loss in publishing data.

The traditional k-anonymity models take all tuples in publishing table as sensitive tuples. So they are to be generalized and the publishing data lost a lot of useful information. The kernel idea of the proposed work is to protect individual's 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. The proposed model not only preserves the privacy of the dataset to a large extent, but also the data utility. The proposed solution guarantees privacy against most of the attacks known to be possible to retrieve private information of individuals. This solution mainly considers the quasi attributes and the attributes which are really sensitive and need to be preserving the privacy of individual are only generalized and anonymized. Other data are published directly. Suppression is avoided to achieve more data utility. It also provides more needed data to researchers and data miners without disturbing the original data.

3. RELATED RESEARCH AND INSIGHTS

K-anonymity alone does not provide full privacy. Suppose attacker knows the non-sensitive attributes (zip, age and nationality) of Chen sui and Jason,

Zip	Age	Nativity	
36900	23	Chinese	← Chen Sui
52013	32	American	← Jason

Fig1. Quasi Data Of Individuals

And the fact that Chinese have very low incidence of heart disease, then Homogeneity and Background knowledge attacks are possible.Consider table 2,

Table 2. Original Data Set

Id	Zipcode	Age	Nativity	Diagnostic
1	42302	25	Indian	Flu
2	52020	22	American	Flu
3	36900	23	Chinese	Heart disease
4	52013	29	American	Cancer
5	42025	31	Indian	HIV
6	13025	38	German	HIV
7	13022	36	German	HIV
8	52013	32	American	HIV

Table2 is anonymized by taking k value as 4, and then the resulting anonymous table looks as in table 3,

Table 3. Anonymous data set (with k=4)

Id	Zipcode	Age	Nativity	Diagnostic	
1	423**	<30	*	Flu	Chen sui Matches here *(Background* knowledge attack)
2	520**	<30	*	Flu	
3	369**	<30	*	Heart disease	
4	520**	<30	*	Cancer	
5	42***	3*	*	HIV	Jason Matches here (Homo - geneity attack)
6	13***	3*	*	HIV	
7	13***	3*	*	HIV	
8	52***	3*	*	HIV	

A privacy threat occurs either when an identity is linked to a record or when an identity is linked to a value on some sensitive attribute. These threats are respectively called record linkage and attribute linkage.

3.1Attacks on k-Anonymous Table

3.1.1 Record Linkage

The record linkage occurs when some values q of quasi-identifiers Q identifies a smaller number of records in the released dataset D. In this case, the record holder having the value q is vulnerable to being linked to a small number of records in D.

The notion of k-anonymity was proposed to fight record linkage. k-anonymity is one of the stronger models of privacy protection. It restricts disclosure threats to an suitable level. The security obtained with k-anonymity is that no information can be linked to groups of less than 'k' individuals. Therefore, the degree of ambiguity of sensitive attribute is at least 1/k. However the main drawback in k-anonymity is its vulnerability to attribute linkage.

3.1.2 Attribute Linkage

If some sensitive values are predominate in a group, an attacker has no difficulty to infer such sensitive values for a record holder belonging to this group. Such attacks are called

attribute linkage. Particularly, k-anonymity suffers from two types of attribute linkage:

- Homogeneity attacks: k-anonymity protection model can create groups that leak information due to lack of diversity in the sensitive attribute. In fact, k-anonymization process is based on generalizing the quasi-identifiers but does not address the sensitive attributes which can reveal information to an attacker.

The table 3 shows the homogeneity attack. It shows how Jason's information is obtained by the homogeneity attack.

- Background knowledge attack:. Beside to homogeneity attacks, the background knowledge attacks can compromise privacy in k-anonymous

database. In fact, an interloper can have knowledge that a priori enables him to guess sensitive data with high confidence.

The table 3 shows the Background knowledge attack. It shows how Jason's information is obtained by the background knowledge attack. This kind of attacks depends on other information available to an attacker.

Using this background knowledge attack, an adversary can disclose information in two ways,

Positive disclosure: In positive disclosure, an adversary can correctly identify the value of a sensitive attributes with high probability.

Negative disclosure: In the negative disclosure the adversary can correctly eliminate some possible values of sensitive attribute with high probability.

So after brief study of these attacks, background knowledge attack is difficult to prevent as compared to homogeneity attack .Given these two weaknesses, several models are introduced to combat attribute linkage. However, the latter are may be difficult to achieve and generally compromise the usefulness and significance of the mining results.

3.2 Multiple Sensitive Attributes

K-anonymity model introduced to protect sensitive attributes from interlopers where 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. Data publisher needs to prevent privacy disclosure which means someone can simply attack link the publish table T and at least know the individuals suffer from some kinds of privacy disease. This phenomenon is a kind of privacy disclosure. Information disclosure is of three types:

Identity disclosure: An individual is linked to a particular record in the published data.

Attribute disclosure: Sensitive attribute information of an individual is disclosed.

Membership Disclosure: Information about whether an individual's record is in the published data or not is disclosed.

So, the data publisher have to convert a private table in such a manner that if an adversary want to search an individual's identity and have knowledge about quasi-identifiers, finds k-1 records that satisfies k-1 quasi-identifiers. Data publishers have to face problem when multiple sensitive attributes are present in records.

Table 4 shows multiple sensitive attributes in adult dataset. In this table Medical Status, Annual income and occupation are considered as a sensitive attributes. So when a data publisher concentrates to protect one sensitive attributes may cause disclosure of identity due to another one. So need a technique to control all sensitive attribute. In the next section the proposed algorithm was explained, which prevent multiple sensitive attributes without suppression.

Table 4. Description of data set

S.NO	ATTRIBUTE	TYPE
1	Zip code	Non- Sensitive
2	Age	Non- Sensitive
3	Nationality	Non- Sensitive
4	Medical Status	Sensitive
5	Occupation	Sensitive
6	Annual Income	Sensitive

4. CONCEPT AND PROBLEM DEFINITION

The objective of proposed work is as follows,

- **Privacy** – To provide the individual data privacy by generalization in such a way that data re-identification cannot 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 Sensitive attribute Based Anonymity Method is to be presented. The key 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{K1,K2.. ,Kj Q1,Q2,..,Qp,S } be a table. For example, T is a medical dataset. Let Q1,…,QP denote the quasi-identifier specified by the application. 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 Kj 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.

Let T be the initial table and T' be the released micro data table. T' consists of a set of tuples over an attribute set. The attributes for k-anonymity table are classified into three categories. Quasi identifier, Key attributes and Sensitive attributes.

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.

Quasi-identifier (QI) attributes are those, such as age and zip code, that in combination, can be linked with external nformation to re-identify (some of) the respondents to whom (some of) the records in the microdata belong. Unlike identifier attributes, QI attributes cannot be removed from the microdata,
because any attribute is potentially a QI attribute.

Definition 2: (Key Attribute):
An attribute K consists of values which is the most unique value for to identify the individual from set S. Denote by K.
Key attributes that can be used to identify a record, such as Name and Social Security Number. Since objective of this paper is to prevent sensitive information from being linked to specific respondents.

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.
Sensitive attributes that are understood to be unknown to the interloper and need to be protected, such as diagnostic report or account number.

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

5. DESIGN PROCESS

5.1 System Architecture

The figure 5.1 clearly outlines every module in the project. The module broadly classifies various sub topics within each of the modules. The input and output of the software forms the boundaries in the given figure.

.

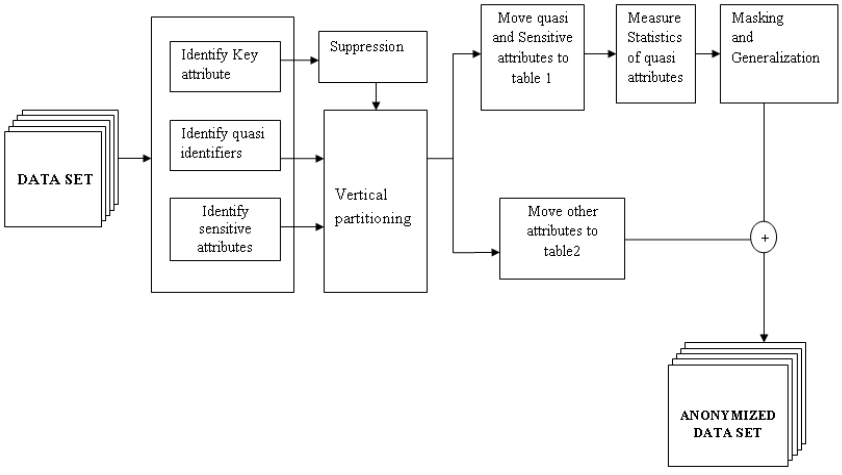

Fig. 2. System Architecture

5.2 Module Description

The proposed work consists of 3 modules

- Identification of attributes
- Vertical partitioning
- Sensitive based Anonymity

5.2.1 Identification of Attributes

The attributes of tables are classified in to three classes. Identify the unique attribute such as id, as a key value. Identify the common attributes that are publicly available in all records as quasi attributes. Sensitive attributes are the attributes which are need to be protected. The selection of sensitive attributes is important because, there is a need to anonymize only the most sensitive data to avoid the overhead and to increase the data utility. After identifying all of the attributes, suppression is applied only to the key attribute.

In this method the key/Identifier attributes have been removed and the quasi-identifier and sensitive attributes are usually kept in the released and initial data set. Basically the values for the sensitive attributes are not available from any external source. These guarantees that an interloper cannot use the sensitive attributes to increase the chances of exposure. Unfortunately, an interloper may use record linkage techniques between quasi-identifier attributes and external available information to gain the identity of individuals from the modified data set. To avoid this possibility of privacy exposure, sensitive attribute based anonymity is anticipated.

5.2.2 Vertical Partitioning

Vertical partitioning divides a table into multiple tables that contain fewer columns. The two types of vertical partitioning are normalization and row splitting. In this step the normalization method was used. Normalization is the standard database process of removing redundant columns from a table and putting them in secondary tables that are linked to the primary table by primary key and foreign key relationships**.** Assign unique class id to the table. After assigning class id divide the table into two. One table containing sensitive data along with quasi identifiers and the other table with non-sensitive data.

Vertical partitioning query scans less data. This increases query performance. For example, a table that contains seven columns of which only the first four are usually referred may help to split last three columns into a separate table. Vertical partitioning must be carefully considered, because analyzing data from various partitions requires query that link the tables. Vertical partitioning also changes the performance if partitions are very large.

5.2.3 Sensitive Based Anonymity

Anonymize only the most Sensitive attributes and quasi attributes to increase the data utility. Masking is applied only to those quasi and sensitive attributes.

The sensitive attributes identified in the previous step is anonymized using different masking techniques. Zip code and credit card number are anonymized using masking technique called shuffling and then age and income are normalized using recoding d to produce the anonymized results.

Then Generalization is applied to the masked data. In order to increase the data utility, suppression is avoided for these attributes.

Generalization is an important technique for protecting privacy in data distribution. In the framework of generalization, k-anonymity is a strong notion of privacy. However, since existing k-anonymity measures are defined in terms of the most specific sensitive attribute (SA) values, algorithms based on these measures can have narrow eligible ranges for data that has a heavily skewed distribution of Sensitive attribute values and produce anonymous data that has a low utility. In this work, generalization is performed in a controlled fashion.

By using the class id as the foreign key join both the sensitive and non-sensitive attribute tables using the sql cross join function to get the anonymized dataset. The result of this cross join enables the dataset to get multiple k-Anonymous records where the impostor finds it difficult to find the exact data of a person.

6. MODEL AND ALGORITHM

6.1 Algorithm

Algorithm for Sensitivity Based Anonymity Method
Input: A dataset D, quasi-identifier attributes Q,
Sensitive values A, Anonymity parameter k
Output: Releasing Table D*

Step 1: Select Data set D from a Database
Step 2: Select Key attribute, Quazi-identifier attribute and Sensitive Attribute from give n attribute list.
Step 3: Select the set of most sensitive values A from list of all sensitive values that is to be preserve.
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 quasi attributes of table T1 i.e. distinct values for that attribute and total no of rows having that value.
Step 6: Apply generalization on quasi identifiers of table T1 to make it k- anonymized.
Step 7: Join both tables T1 and T2.
T*=T1+T2 which is table ready to release.

7. EXPERIMENTAL RESULTS

This method is computed on the Bank Dataset from the UCI Machine Learning Repository. The Bank Dataset contains 32561 tuple. After preprocessing data and removing tuples containing missing values 30162 tuples are selected. This database contains 17 attributes from that only 4 attributes are considered as sensitive.

Table 5 provides a brief description of the data including the attributes used in method, the number of distinct values for each attribute, the generalization that was used for Quazi identifier attributes and the height of the generalization hierarchy for each attribute.

Table 5 Description of Bank 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	

Figure 3 shows privacy vs. utility graph for the sensitive attribute based anonymity. From figure it is clear that the traditional k-anonymity algorithm yields more information loss because all attributes in the data set are generalized. Only sensitive attributes are generalized in this approach.

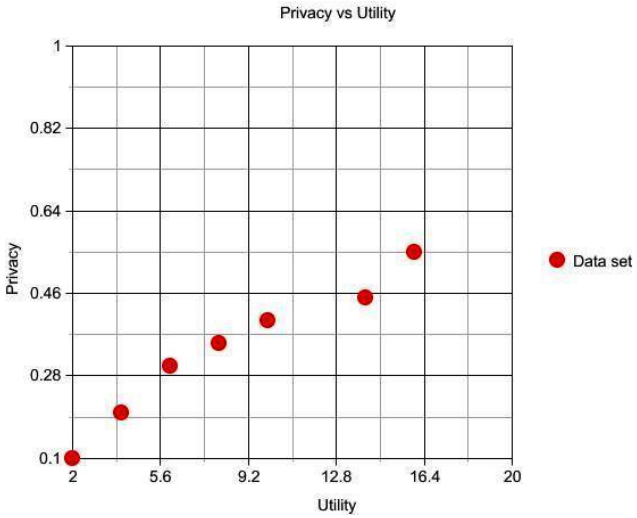

Fig 3. Privacy vs. Utility

The significance of this algorithm is not limited to theory only. Experiments on several databases using several measures of information loss show that the proposed algorithm constantly achieves lower information losses than the currently using algorithms. Hence, in terms of achieving minimal information losses, this algorithm appears to offer the best performance. Further, the runtime of the algorithm is reduced, due to the anonymization of selective attributes.

8. 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

So there is a need to develop a method which provides the privacy with minimum information loss and maximum data utility. This paper present a new k-anonymity model based on sensitive attributes, by which information loss is reduced. Only sensitive attributes are anonymized by this method, so data utility is increased. Excessive generalization and suppression leads to the reduction of data utility and more information loss of publishing data. A Sensitive attribute based Anonymity Method generalizes the high sensitive value alone along with quasi attributes. This method avoids suppression to increase data utility. When suppression is applied, record are ignored which causes data loss. This is a secure algorithm to maintain usability and privacy of data sets.

9. REFERENCES

[1] G. Loukides, A. Gkoulalas-Divanis, "Utility-preserving transaction data anonymization with low information loss", Expert Systems with Applications, Elsevier 2012.

[2] Batya Kenig and Tamir Tassa "A practical approximation algorithm for optimal *k-anonymity*", Data Mining Knowledge Discovery,Springer,2011

[3] Pingshui WANG, "Survey on Privacy Preserving Data Mining", International Journal of Digital Content Technology and its Applications, December 2010.

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

[5] John Miller, Alina Campan and Traian Marius Truta, "Constrained k- Anonymity: Privacy with Generalization Boundaries ", The VLDB Journal-The International Journal on Very Large Databases ,2008.

[6] 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

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

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

[9] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy" , International Journal on Uncertainty Fuzziness Knowledge based Systems,2002

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

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