Multiple Sensitive Attributes based Privacy Preserving Data Mining using k-Anonymity

P.Usha, R.Shriram, S.Sathishkumar

Abstract— Data mining is the process of extracting interesting patterns or knowledge from large amount of data. With the development of Data mining technology, an increasing number of data can be mined out to reveal some potential information about the user, because of which privacy of the user may be violated easily. Privacy Preserving Data Mining is used to mine the potential valuable knowledge without revealing the personal information of the individuals. k-anonymity is one of the Privacy Preserving model that aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The existing approaches are based on homogeneous anonymization that anonymizes quasi attributes by choosing a single sensitive attribute. This approach causes high information loss and reduces the data utility. To overcome these issues in the existing system, Clustering based non-homogeneous anonymization system is proposed. In the proposed system, instead of selecting a single attribute, multiple sensitive attributes are selected. Generalization technique is applied on the most sensitive attribute and it is clustered. Based on the sensitivity level of the clusters, non-homogeneous anonymization technique (generalization and suppression) is applied to the identified quasi attributes of each cluster. The remaining non sensitive attributes are directly published. Thus the proposed system achieves high degree of data utility, reduces information loss and also achieves high degree of Data Integrity.

Index Terms— Data mining, Generalization, Homogeneous anonymization, k-Anonymity, Non-homogeneous anonymization, Privacy Preserving Data Mining (PPDM), Suppression.

———————————— ◆ ————————————

1 Introduction

With the increased digitization of the world, more and more information about the individuals is collected by governments and corporations and stored in various databases. The collection of information has created massive opportunities for knowledge based decision making. Driven by either mutual benefits or by regulations of public available information, there is a demand for the exchange and publication of data among various parties. As a result, there is an enormous quantity of privately owned records that describe individual's finances, interests, activities and demographics. These records often include sensitive data and may violate the privacy of the individual's if published. This information is becoming a very important resource for many systems and corporations that may enhance their services and performance by inducing novel and potentially useful data mining models. One common practice for releasing such confidential data without violating privacy is applying regulations, policies and guiding principles for the use of the data. Such regulations usually entail data distortion operations such as generalization or random perturbations. The major challenges in this approach are data leakage and ineffectiveness of resultant data due to excessive data distortion.

The emerging research field in data mining, Privacy Preserving Data Publishing (PPDP) is targeting these challenges. It aims at developing techniques that enable publishing data while minimizing data distortion for maintaining utility and ensuring that privacy is preserved. In this paper a new privacy preserving data publishing method is proposed, which is called clustering based non-homogeneous anonymization algorithm.

The attributes in the database table is distinguished into four types that needs to be published namely Key identifiers, attributes that uniquely identify an individual (e.g. ssn, name);Quasi-identifiers, publicly-accessible attributes that do not identify a person, but some combinations of their values might yield unique identification (e.g., city, gender, age, and zipcode); Sensitive attributes, private information of individual's such as medical or financial data; and Other non-sensitive attributes that, on one hand, cannot be used for identification since they are unlikely to be accessible to the adversary, and do not represent information of sensitive nature. (Those attributes can be ignored in our discussion.) A common practice in PPDP and PPDM is to remove the key identifiers and to generalize or suppress the quasi-identifiers in order to protect the sensitive data of individuals from being revealed. In case of generalization the original values of quasi-identifiers are replaced with less precise values and in case of suppression no values are released at all. The sensitive data is usually retained unchanged. In the past years, several models were suggested for maintaining privacy when disseminating data. Most approaches evolved from the basic model of k-anonymity. In that model, the practice is to remove the identifiers and generalize the quasi-identifiers as described above, until each generalized record is indistinguishable from at least k-1 other generalized records based on sensitive attribute. Consequently,

- *P.Usha is a Research Scholar in School of Computer, Information and Mathematical Sciences, B.S.Abdur Rahman University, Chennai. E-mail: usha@bsauniv.ac.in*
- *Dr.R.Shriram is currently working as Professor in School of Computer, Information and Mathematical Sciences, B.S.Abdur Rahman University, Chennai. E-mail: shriram@bsauniv.ac.in*
- *S.Sathish kumar is currently pursuing master's degree program in School of Computer, Information and Mathematical Sciences, B.S.Abdur Rahman University, Chennai. E-mail: sathish.thangam@yahoo.com*

an adversary who wishes to trace a record of a specific person in the anonymized table will not be able to trace that person's record to subsets of less than k anonymized records.

Table 1:Attribute Classification

S.No	Attribute	Type
1	ID	Key
2	Zipcode	Quasi
3	Age	Quasi
4	Nationality	Quasi
5	Disease	Sensitive
6	Salary	Sensitive

As an example, consider the basic table in Table 1, having the key attribute-ID; quasi-identifiers-Age, Nationality and Zipcode; and the sensitive attributes- Disease and Salary.

2 Previous Related Work

Latanya sweeney [12] presented a model named k-anonymity and a set of accompanying policies for deployment. The released dataset of the model provides k-anonymity protection if 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. This paper also examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, m-Argus and k-Similar provides guarantee of privacy protection. But it does not prevent the record and attribute linkage attacks.

Jiuyong Li, Raymond Chi-Wing Wong, AdaWai-Chee Fu, and Jian Pei [11] proposed a model to achieve k-anonymity by clustering in attribute hierarchical structures. Authors have defined generalization distances between tuples to characterize distortions by Generalizations and discuss the properties of the distances and concluded that the Generalization distance is a metric distance. This paper proposes an efficient clustering based algorithm for k-anonymization and experimentally shows that the proposed method is more scalable and causes significantly less distortions than an optimal global recoding k-anonymity method.

A. Machanavajjhala [10] presented the model to solve the linkage attacks in k-anonymity called l-diversity. In addition to building a formal foundation for l-diversity, experimental evaluation of l-diversity is practical and can be implemented efficiently. It tries to put constraints on minimum number of distinct values seen within a equivalence class for any sensitive attribute. Even though this method prevents the linkage attacks, it suffers from homogeneity and background knowledge attacks.

Arik Friedman, Ran Wolff, Assaf Schuster [9] presented extended definitions of k-anonymity and used them to prove that a given data mining model does not violate the k-anonymity of the individuals represented in the learning examples. It also describes data mining algorithms that generate only k-anonymous output. Finally, this method contributes new and efficient ways to anonymize data and preserve patterns during anonymization.

Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach [8] proposed a new method of using k-anonymity for preserving privacy in classification tasks Instead of suppression they proposed swapping which decreases information loss induced by the suppression approach. The new method also shows a higher predictive performance and less information loss when compared to existing state of-the-art methods.

Pingshui WANG [6] presented a wide survey of different privacy preserving data mining algorithms and analyses the representative techniques for privacy preserving data mining, and points out their merits and demerits. They also discuss present problems and directions for future research.

Batya Kenig, Tamir Tassa [5] proposed modified k-anonymity method. The process of anonymizing a database table typically involves generalizing table entries and, consequently, it incurs loss of relevant information. The modified algorithm that issues l-diverse k-anonymizations also achieves lower information losses than the corresponding modified versions of the leading algorithms. Experiments show that the proposed algorithm provides smaller information losses than the best known approximation algorithm as well as the best known heuristic algorithms.

Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy [4] proposed a novel technique called slicing, which partitions the data both horizontally and vertically. The slicing algorithm preserves better data utility than generalization, more effective than bucketization in workloads involving the sensitive attribute, the sliced table can be computed efficiently and shows the effectiveness of slicing in membership disclosure protection.

Junqiang Liu [3] provides a comparative analysis of the state of the art works along multiple dimensions. Privacy Preserving Data Publishing research is motivated by real world problems which however are far from being solved as there are still challenging issues to be addressed. This study helps to identify challenges, focus on research efforts and highlight the future directions.

G. Loukides, A. Gkoulalas-Divanis Liu [1] proposed a novel approach for anonymizing data in a way that satisfies data publisher’s utility requirements and incurs low information loss. To achieve this, they introduced an accurate information loss measure and an effective anonymization algorithm that explores a large part of the problem space.

Traditional k-anonymity models consider single attribute as sensitive attribute and perform homogeneous anonymization for all remaining attributes. There is loss of useful information and reduction in data utility, while using homogeneous anonymization. To overcome the issues in the existing models, Clustering based non-homogeneous anonymization system is proposed to achieve high degree of data utility and data integrity by reducing information loss.

3 Concept and Problem Definition

The objective of this work is to provide privacy to the individuals data by generalization in such a way that data re-identification cannot be possible. 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 Clustering based non-homogeneous anonymization. In this system, instead of selecting a single sensitive attribute, multiple sensitive attributes are selected. Generalization technique is applied on the most sensitive attribute and it is clustered. Based on the sensitivity level of the clusters, non-homogeneous anonymization technique (generalization and suppression) is applied to the identified quasi attributes of each cluster. The remaining non sensitive attributes are directly published.

3.1 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, Q_2……, Q_P 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 namely quasi identifiers, Key attribute and Sensitive attributes.

3.2 Key Attribute

An attribute denoted by ‘K’ consists of values which is the most unique value to identify the individual from dataset ‘S’. Key attributes are used to identify a record, such as Name and Social Security Number.

3.3 Quasi Identifiers

A set of non-sensitive attributes {Q_1, Q_2, …,Q_p} of a table is called a quasi-identifier, if these attributes can uniquely identify (can be called as candidate key) at least one individual in the general population when linked with external data. Quasi-identifier (QI) attributes are those, such as age and zip code.

3.4 Sensitive Attributes

A set ‘A’ consists of values which the user selects as most sensitive attributes from dataset ‘S’. These attributes is what the researchers need, so they are always released directly. Sensitive attributes such as medical data (disease), salary, account number, etc. that are understood to be unknown to the interloper needs to be protected.

4 System Architecture

The system architecture is shown in Figure 1 that clearly outlines every module. 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. This work consists of five main modules such as 1.Preprocessing Dataset, 2.Identification of attributes, 3.Vertical partitioning, 4.Sensitive attribute generalization and 5.Clustering based non-homogeneous anonymization.

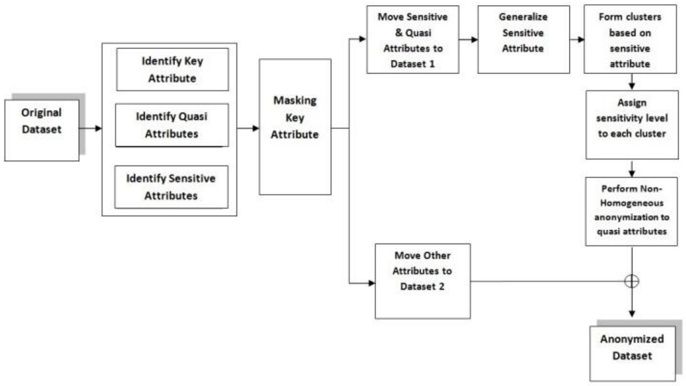
Original Dataset
Identify Key Attribute
Identify Quasi Attributes
Identify Sensitive Attributes
Masking Key Attribute
Move Sensitive & Quasi Attributes to Dataset 1
Generalize Sensitive Attribute
Form clusters based on sensitive attribute
Assign sensitivity level to each cluster
Perform Non-Homogeneous anonymization to quasi attributes
Move Other Attributes to Dataset 2
Anonymized Dataset

Fig.1: Clustering based non-homogeneous anonymization

The Database or dataset is collected from different organizations like insurance agencies, hospitals, etc. (E.g. Insurance Dataset, Medical Dataset).The missing attributes or data values in the dataset are removed and then converted into csv(comma separated files)files or text files for processing.

4.2 Identification of Attributes

The attributes in database which comes under four categories namely Key attribute, Quasi attribute, Sensitive attribute and non-sensitive attributes. The key (or) Identity attribute, Quasi attributes and Sensitive attributes are modified by using generalization and suppression methods to preserve the privacy of the individuals. The non-sensitive attributes in dataset are directly published to maintain the data utility.

The key or unique attribute is identified from the dataset such as id, ssn (social security number), name that uniquely identify the individuals. The attributes that are publically available are identified as quasi attributes from the dataset such as age, zip code, date of birth, gender. Identify the sensitive attributes that contains individuals private information such as medical data (diagnostics), financial data (salary), etc. In this model, the key identifier is removed totally which directly identifies the individuals. The quasi attributes are generalized or suppressed until each generalized record is indistinguishable from at least k-1 other generalized records.

4.3 Vertical Partitioning

Vertical partitioning divides the database table into multiple tables that contain fewer columns. 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 is performed on the table to split into two, of which one containing sensitive data along with quasi attributes (Dataset 1) and other containing non sensitive attributes (Dataset 2).

4.4 Sensitive Attribute Generalization

In this approach multiple attributes are selected as sensitive attributes in order to provide high degree of data utility. From the selected attributes, one attribute is chosen as clustering attribute and generalization is performed on that attribute [16]. For example if we choose disease and profession as sensitive attributes. An example for sensitive attribute generalization is shown in Table 2. E.g. Disease

Table 2: Generalization of Sensitive attribute

Symptoms/Disease	Disease groups
Diabetes	Kidney
Kidney	
Urination Problems	Heart
Palpitation	
Heart	Lungs
Angina	
Dry Skin	

4.5 Clustering based non-homogeneous anonymization

The generalized sensitive attribute is clustered. Different anonymization rule is applied for each cluster. Based on the sensitivity level of the clusters corresponding quasi attributes are anonymized using non-homogeneous anonymization technique. An example for homogeneous anonymization of the medical dataset is shown in Table 3.

Table 3. Homogeneous Anonymization

Age	Gender	Zip-code	Income	Marital Status	Disease
[25-45]	Male	1****	9000	Single	Flu
[25-45]	Male	1****	35000	Divorced	AIDS
[25-45]	Female	1****	18000	Married	Cancer
[25-45]	Male	1****	9000	Married	Flu

Table 4. Non-Homogeneous Anonymization

Age	Gender	Zip-code	Income	Marital Status	Disease
28	Male	130**	9000	Single	Flu
[25-45]	*	1****	35000	Divorced	AIDS
[25-45]	Female	13***	18000	Married	Cancer
31	Male	105**	9000	Married	Flu

An example for Non-homogeneous anonymization is shown in Table 4. Finally, the anonymized quasi and sensitive dataset is joined with Dataset 2 that contains non-sensitive attributes. The result of this join produces anonymized dataset of original dataset with high degree of data utility and reduction in information loss.

5 METHODOLOGY

The following algorithm shows step by step procedure of this system.

```
Input: A Dataset or Table T [Key attribute K, Quasi
       Attributes Q, Sensitive attributes S, Non-sensitive
       attributes A]
Output: Anonymized Dataset T*.

Begin
  Select Dataset T
  Identify Key attribute, Quasi attributes and Sensitive attributes from then Table T.
  for each tuple in Dataset T do
     Suppress key attribute $K_i$
  end for
  for each tuple in Dataset T do
     Quasi attributes Q and sensitive attributes S are moved to Table T1
     Remaining attributes or Non-sensitive attributes A are moved to Table T2
  end for
  for each tuple in Table T1 do
     Generalize the most sensitive attribute
     Form clusters based on generalized sensitive attribute
     Assign sensitivity level  for each cluster
     Apply non-homogeneous anonymization on quasi attributes based on sensitivity level of the cluster to make it k-Anonymized
  end for
  Join Table T1 and Table T2.T* = T1+ T2
  Let T* be the anonymized Dataset
End
```

6 EXPERIMENTAL ANALYSIS

The experimental study was performed on a Medical Dataset comprising of 500 records with 12 attributes which includes key attribute, 5 quasi attributes, 2 sensitive attributes and 4 non-sensitive attributes.

Table 5. Group based Anonymization

Attribute Name	No. of Distinct records	No. of groups in HA	No. of groups in CNHA
City	47	2	2
Age	57	7	35
Zipcode	487	3	24
Gender	2	2	3
Diagnosis	113	Nil	7

Table 6. Comparison of HA and CNHA

Attribute Name	Disclosure Rate		Privacy Rate	
	HA	CNHA	HA	CNHA
Age	1/7=14.28	1/35=2.86	85.72	97.14
Gender	1/2=50	1/3=33.33	50	66.67
City	1/2=50	1/2=50	50	50
Zipcode	1/3=33.33	1/24=4.16	66.67	95.84
Diagnosis	1/1=100	1/7=14.28	0	85.72

Table 5. describes the number of distinct records and the number of groups formed by Homogeneous Anonymization (HA) and Clustering based Non-Homogeneous Anonymization (CNHA).

Table 6. gives a comparison of HA and CNHA based on their disclosure rate and privacy rate. Based on Table 6. the following graphs are drawn which depicts decreased disclosure rate and increased privacy rate in this proposed system (CNHA).

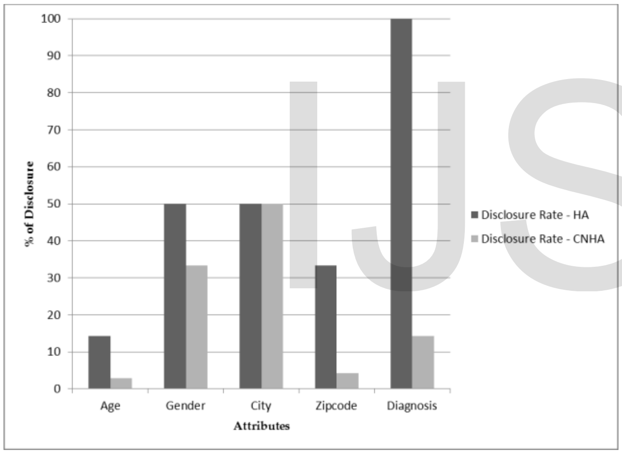

Fig. 2: Disclosure Rate based comparison of HA and CNHA

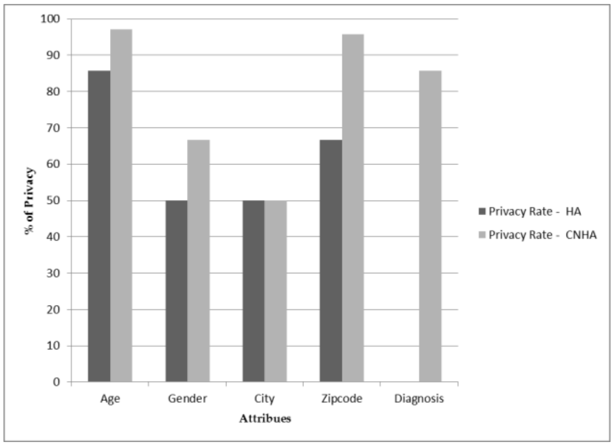

Fig. 3: Privacy Rate based comparison of HA and CNHA

7 Conclusion

In this paper a new privacy-preserving data publishing algorithm called clustering based non-homogeneous anonymization algorithm is presented that anonymizes the quasi-identifiers based on the sensitivity level of the cluster. This technique performs non-homogeneous anonymizations, which reduces information losses and provides high degree of data utility than the data mining tasks that involves homogeneous anonymizations. This system achieves a minimum disclosure rate of 2.86% and a maximum privacy rate of 97.94%.

References

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

[2]. Bhavana Abad (Khivsara) and Kinariwala S.A , “A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity”, *International Journal of Computer Applications*, March 2012.

[3]. Junqiang Liu , " Privacy Preserving Data Publishing: Current Status and New Directions", *Information Technology Journal 11(1):1-8*,2012

[4]. Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy , " Slicing: A New Approach for Privacy Preserving Data Publishing", *IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3*, March 2012

[5]. Batya Kenig · Tamir Tassa , " A practical approximation algorithm for optimal *k*-anonymity", *Springer Data Mining Knowledge Discovery* (2012) 25:134–168

[6]. Pingshui WANG, “Survey on Privacy Preserving Data Mining”, *International Journal of Digital Content Technology and its Applications*, December 2010.

[7]. Charu Aggarwal, Philip Yu, “Models and Algorithms: Privacy-Preserving Data Mining”, *Springer* 2008.

[8]. Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach , “KACTUS 2: Privacy Preserving in Classification Tasks Using k-Anonymity”, *Springer-Verlag Berlin Heidelberg* 2009

[9]. Arik Friedman, Ran Wolff, Assaf Schuster , “Providing k-Anonymity in Data Mining”, *The VLDB Journal - The International Journal on Very Large Data Bases archive Volume 17 Issue 4*, July 2008

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

[11]. Jiuyong Li, Raymond Chi-Wing Wong, AdaWai-Chee Fu, and Jian Pei, " Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures", *Springer-Verlag Berlin Heidelberg* 2006.

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

[13]. P.Samarati, “Protecting respondents identities in microdata release”, *IEEE Transactions on Knowledge and Data Engineering, 13(6):10101027*.2001.

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

[15]. X-H Zhang, Z L Lu, L Liu, “ Coronary heart disease in China “, Global burden of cardiovascular disease, July 20, 2012.

[16]. Health and Social Care Information Centre http://www.hscic.gov.uk/

[17]. National Casemix Office http://www.hscic.gov.uk/casemix