SCBA: A New Sensitivity Based Anonymity Method for Privacy Preserving in Medical Data Mining

Bhavana V. Abad(Khivsara) **S. A. Kinariwala**

ABSTRACT **- For the data holder, such as a medical institution, public health agency, share person, record data in such a way that the released information remain practically useful, but the identity of the individuals can be also be anonymized. Here the privacy preserving in data mining comes into picture. Many medical applications employing the data mining techniques involve mining the data that includes private and sensitive information about the subjects. k-anonymity is a property that models the protection of released data against possible re-identification of the respondents to which the data refers. Traditional k-anonymity, however has the problem of information loss and data utility as it considers all tuples of publishing database as sensitive and so all tuples are involved in anonymity, which lead to reduce the precision of publishing table. In this paper, we present a new method called Sensitivity Class Based Anonymity which solves this problem by classifying the tuples into two classes high and low, according to their sensitive value. Only tuples from high class are involved into generalization, and the low class tuples are released as it is.**

Keywords
Data Mining, Privacy, k-Anonymity, re-identification

I. INTRODUCTION

There are number of huge databases available today in the society, like medical patient data, census data, media related data and data collected by different government agencies. These huge databases can be utilized for medical research purpose, to improve customer service or for home land security purposes. Many organizations collect and hold these large volumes. They would like to publish the data for the purposes of data mining which is useful in many domains for decision making. But the problem with data mining output is that, it also reveals some information which is considered to be private and sensitive, so the privacy is becoming very important in many data mining applications.[2]

Following example shows how the released data can be used to re-identify the individual. Suppose, the medical data set as shown in Table I is published by data miner and Table II is any data base available publically like voter list. By linking these two tables, the attacker can easily re-identify that,Arjun is suffering from cancer and in this way the privacy of individual is violated. This is happened because the combination of values of Quazi attributes like Zip code, Age and Sex is unique in medical data set.

Table I Medical Data set

ID	ZIPCODE	AGE	SEX	DIAGNOSIS
1	423065	29	M	Heart Disease
2	422036	32	F	Flu
3	423245	38	M	Headache
4	422035	27	F	HIV
5	423012	47	M	Cancer
6	423432	53	F	Viral

Table II Voter List

NAME	ZIPCODE	AGE	SEX
Mohit	423065	29	M
Sunil	422036	32	F
Shyam	423245	38	M
Rohini	422035	27	F
Arjun	423012	47	M
Sangita	423432	53	F

Here, the problem is how to preserve the privacy of the individual. The solution to this is by releasing the data set in such a way that there is no unique combination for quazi-identifiers.

II. PRIVACY PRESERVING APPROACHES

In order to protect the privacy of the individual, data is manipulated or processed in such a way that private data remains private even after the mining process. These are various methods used for Privacy Preserving in Data Mining are as follows [1]:

1. Statistical Methods :
   - Randomization methods
   - Swapping
   - Micro Aggregation
   - Synthetic data generation
2. Group based anonymization methods:
   - k-anonymity
   - l-diversity
   - t-closeness
3. Personalized privacy preserving
4. Utility based privacy preserving
5. Distributed privacy preserving data mining using cryptographic method
   - Horizontal partitioning
   - Vertical partitioning

III. APPROACHES BASED ON ANONYMITY BASED MODEL

In anonymity based method, data provider often removes key attributes such as names, addresses, phone number. De-identifying data, however, provides no guarantee of anonymity. Released information often contains other data called as quasi identifiers such as, birth date, sex, and ZIP code, 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, 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 as the heuristics determined from the transformed data are close to original heuristics and the privacy of users is not dying out.

Definition 1 (Micro data): The data to be released after applying anonymization methods is called the Micro Data.

Definition 2 (Sensitive attribute): Attribute which must not be disclosed in the released micro data.

Definition 3 (Quasi Identifier) : Attributes or combination of attributes within a dataset which on their own are non-sensitive but on combination with external data are capable of identifying records.

Definition 4 (Equivalence Class): All set of tuples which cannot be distinguished from each other with respect to Quasi-Identifier are called an Equivalence Class.

A. k- ANONYMITY

In 2002 Sweeney [2] proposed a method that achieves 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 other words, we should not be able to make ANY query to the database which returns less than 'k' matches. In k-anonymity the attributes of tables are classified in three classes as shown in Table III. First is key attribute which is generally the primary key like name and it is removed at time of releasing, second class is quasi-identifier that are generally linked with publicly available database to re-identify the individual like age, zip code, gender, birth date and these attributes are generally suppressed or generalized . Third class is sensitive attribute used by researchers and generally released directly [4].

Table III Classification of Attributes for k-anonymity

Key attribute	Quasi_identifier			Sensitive attributes
Name	Gender	Age	Zip code	Diagnosis
John	Male	25	423101	depression
Smith	Male	27	423109	HIV
Bob	Male	22	423508	Flu
Jenna	Female	43	425221	Cancer

k-anonymity is provided by use of generalization relationships between domains and between values that attributes can assume. Suppression is a complementary approach to providing k-anonymity.

Definition 5 (Generalization): Given two domains D1 and D2 , $D1 \leq D2$ describes the fact that values of attributes in D2 are more generalized values.

Definition 6 (Suppression): Removing data(ie. rows) from table so that it is not released in the micro data is called suppression.

Definition 7 (K-minimal Generalization with suppression): Generalization T1 is k- minimal if it satisfies k-anonymity, it does not enforce more suppression than allowed(some predefined parameter), and there does not exist another generalization satisfying these conditions less general than T1 .

Definition 8 (Identity disclosure): An individual can be directly linked to a particular record in the released data.

Definition 9 (Attribute disclosure): The chances of guessing the sensitive attribute of an individual increase because of released micro data .

K-anonymity : Attacks

k-anonymity cannot provide a safeguard against attribute disclosure in all cases. A simple case of attribute disclosure will be when all the sensitive attributes within a equivalence class have same value. Here we would have achieved k-anonymity but we can accurately predict the sensitive attribute of any person who we can match to this equivalence class by using information in public domain[5].k-Anonymity does not provide privacy if sensitive values in an equivalence class lack diversity.

The two tables show the original and anonymous version of the dataset. In the second table we have 3 equivalent classes. We have achieved 3-anonymity by generalization. The Disease attribute is sensitive. Let us assume that,the attacker can get from public information Sangita's age, say 34 and Zip code, say 27678. Alice also knows that Sangita's record is among one of the records in the original table. From second table, attacker can figure out that Sangita's record is from first equivalence class and can thus figure out that she is suffering from Heart Disease. This attack called as homogeneity attack, which predicts the sensitive attribute because the sensitive values in an equivalence class lack diversity.

Voter List

Sangita	
Zip code	Age
27678	34

3-anonymous patient table

Zip Code	Age	Disease
27***	3*	Heart Disease
27***	3*	Heart Disease
27***	3*	Heart Disease
345**	2*	Flu
345**	2*	Cancer
345**	2*	Cancer

Fig 1 Homogeneity Attack

B. l-DIVERSITY

In 2006 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. An equivalence class has *l* -diversity if there are*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.

Table IV 3-diverse patient table

Zip code	Age	Disease
276**	3*	Gastric Ulcer
276**	3*	Gastric
276**	3*	Stomach Cancer
478**	40	Flu
478**	40	Bronchitis
478**	40	Gastric
479**	2*	Bronchitis
479**	2*	pneumonia
479**	2*	Stomach cancer

As shown in table IV Each Equivalent class has different values in sensitive attribute field. But still there are some limitations of l-diversity that are

1. l-diversity is unnecessary and difficult to achieve for some cases
2. Skewness Attack- Although identity disclosure is successfully handled by l-diversity, but it does not prevent attribute disclosure when the overall distribution is skewed.
3. l-diversity is difficult to achieve
4. l-diversity does not consider the overall distribution of sensitive values

C. t-CLOSENESS

S. Venkatasubramanian in 2007 [3] present a model called t-closeness model that was introduced to overcome attacks possible on l-diversity like similarity attack. l-diversity model uses all values of a given attribute in a similar way even if they are semantically related. Also not all values of an attribute are equally sensitive.

The t-closeness 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 .Now a distance metric between 2 distributions is desired .Thus Earth Mover's distance (EMD) is used. The EMD is based on the minimal amount of work which has to be done to transform one distribution to another by moving distribution mass between each other.

Limitations oft –closeness Anonymity Model

1. There is no computational procedure to enforce t-closeness.
2. Lost co-relation between different attributes: This is because each attribute is generalized separately and so we loose their dependence on each other.
3. Utility of data is damaged if we use very small t.

D. (k,p,q,r)ANONYMITY MODEL

(k,p,q,r)-anonymity[9]is said to be achieved if the data satisfies p-sensitivity for groups where confidential attributes appear very less frequently(less frequent than parameter q).For such groups(after p-sensitivity constraint),the ratio of variance with in the group of sensitive attributes and variance within entire data set is at least r.

Limitations of(k,p,q,r)Anonymity Model

1. It does not provide any defense against skewness attack which was one of the main reasons for using distance metric like EMD in t-closeness model.

E. INCOGNITO: EFFICIENT FULL DOMAIN K-ANONYMITY

K. LeFevre, David J. ,R. Ramakrishnan[7] describes that, they provide a practical framework for implementing one model of anonymization, called full-domain generalization. They introduce a set of algorithms for producing minimal full-domain generalizations, and show that these algorithms perform up to an order of magnitude faster than previous algorithms on two real-life databases.

The Incognito algorithm generates the set of all possible k-anonymous full-domain generalizations of *T*, with an optional tuple suppression threshold. Based on the subset property, the algorithm begins by checking single-attribute subsets of the quasi-identifier, and then iterates, checking k-anonymity with respect to increasingly large subsets. Each iteration consists of two main parts :

1. Each iteration considers a graph of candidate multi-attribute generalizations (nodes) constructed from a subset of the quasi-identifier of size *i*. We denote the set of candidate nodes *Ci*. The set of direct multi-attribute generalization relationships (edges) connecting these nodes is denoted *Ei*. A modified breadth-first search over the graph yields the set of multi-attribute generalizations of size *i* with respect to which *T* is k- anonymous (denoted *Si*).
2. After obtaining *Si* , the algorithm constructs the set of candidate nodes of size $i + 1$ ($Ci+1$), and the edges connecting them ($Ei+1$) using the subset property.

F. (α, k)-ANONYMITY

R. Wong, J. Li, A. Fu, K. Wang [8] propose an (α, k)-anonymity model to protect both identifications and relationships to sensitive information in data. They discuss the properties of (α, k)-anonymity model. They prove that the optimal (α, k) -anonymity problem is NP-hard. They present an optimal global recoding method for the (α, k)-anonymity problem. Next they propose a local-recoding algorithm which is more scalable and result in less data distortion. The effectiveness and efficiency are shown by experiments. Also describes how the model can be extended to more general cases.

IV. SCBA: A NEW PROPOSED METHOD

A survey of the broad areas of privacy-preserving data mining and the underlying algorithms has been done. Also a variety of k-anonymity based methods are also reviewed, so there is a need to develop a method which provides the privacy with minimum information loss and maximum data utility.

As k-anonymity method considers all tuples as equally sensitive and so all tuples get anonymized which leads to more information loss. To solve this problem we develop a novel method called sensitivity class based anonymity(SCBA) that overcomes all above problems. The core of our solution is the concept of classifying the sensitive attribute values in two categories:

1.High Sensitive values – A set of sensitive attribute values H= {s1, s2…sn} that are highly sensitive like HIV , Cancer.
2.Low Sensitive Values – A set of sensitive attribute values L= {s1, s2…sk) that are low sensitive.

SCBA Algorithm

Input – Table T , set of Quazi identifier Q,
Output – Anonymized table T*

Step1: Select Input table and Qset of quazi-identifier attributes
Step2: Select sensitive attribute S.
Step3: Classify sensitive values in two classes H and L
Step4: For each tuple whose sensitive value belongs to set H i.e. if t[S] ∈ H then move all these tuples to Table T1,and apply generalization on Quazi attribute so that tuples get anonymized
Step 5: If t[S] ∈ L then move all these tuples to Table T2
Step 7: Append rows of table T1,T2.
T*=T1+T2 which is table ready to release.

Table Vis the outcome table, after applying traditional k-anonymity and table VI is the T* table after applying SCBA algorithm on Table I. Sensitive values HIV and Cancer are selected as High sensitive value and tuples belonging to that values are moved to table T1 and generalization is applied on quazi attributes Zipcode, Age and Sex to anonimize those tuples. Sensitive values like Flu and Headache etc are selected as Low sensitive values and they are released as it is.

By comparing these two outputs we can see that in traditional k-anonymity the information loss is more as compared to SCBA.

Table V: After applying Traditional k-anonymity on Table I

ZIPCODE	AGE	SEX	DIAGNOSIS
42****	30	M	Heart Disease
42****	30	F	Flu
42****	30	M	Headache
42****	30	F	HIV
423***	>45	*	Cancer
423***	>45	*	Viral

Table VI After applying SCBA on Table I

ZIPCODE	AGE	SEX	DIAGNOSIS
423065	29	M	Heart Disease
422036	32	F	Flu
423245	38	M	Headache
42****	>25	*	HIV
42****	>25	*	Cancer
423432	53	F	Viral

V. EXPERIMENTAL RESULTS

This method is computed on the Adult Database from the UCI Machine Learning Repository [10]. The Adult Database from US Census data. After preprocessing data and removing tuples containing missing values 30162 tuples are selected. This database contains 11 attributes from that only 4 attributes are used. From that four attributes we consider 'occupation' as a sensitive attribute.

Table VII 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 VII Description of Adults Data Set

	Attribute	Distinct	Generalizatio	Height
1	Age	74	10-,20-,30	4
2	Marital	7	Taxonomy	3
3	Race	5	Taxonomy	2
4	Sex	2	Suppression	1
5	Occupation	14	Sensitive	

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_{ij} denote the height of the generalization hierarchy for tuple t_j in attribute Ai.

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

VI. CONCLUSION

While k-anonymity protects against the identity disclosure, it does not provide protection against homogeneity attack and background knowledge attack**.** The *l*-diversity attempts to solve this problem. But *l*-diversity itself has certain limitations. To prevent this *t*-closeness is attempted. All these models try to solve the problem of privacy disclosure with cost of information loss. Hence, a new method is proposed which provides the privacy to individual with minimum information loss and maximum data utility.

REFERENCES

[1] CharuAggarwal , Philip Yu " Models and Algorithms : Privacy-Preserving Data Mining" *Springer* 2008

[2] L. Sweeney. k-Anonymity: "A Model for Protecting Privacy" I*nternational 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 and l-Diversity". *In: Proceedings of the IEEE ICDE (2007)*

[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] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy,DThomas and A. Zhu. "Anonymizing tables", *In Proc. of the 10th International Conference on Database Theory(ICDT05)*, pp. 246-258Edinburgh, Scotland. 2005

[7] 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

[8] R. Wong, J. Li, A. Fu, K. Wang." (α, k)-anonymity: an enhanced k-anonymity model For privacy preserving data publishing", *In: Proceedings of the KDD 2006*:754-759

[9] Josep Domingo-Ferrer, FrancescSebe, AgustiSolanas An Anonymity Model Achievable Via Microaggregation (Secure Data Management 2008)

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

[11] X.H. ,J.D. 2009 "K-Anonymity Based on Sensitive Tuples"*In: Proceedings of the IEEE First International Workshop on Database Technology and Applications 2009*

AUTHOR'S PROFILE

BhavanaAbad(Khivsara) obtained her Bachelors degree in Computers from BAMU university during 2001 and currently pursuing Masters Degree from the same university. She is working in Computer Dept of SNJB's COE ,Chandwad,Pune University as Lecture from last 5 years. She is the member of CSI . Her reseach area includes Privacy Preserving in Data Mining. She has published paper on "A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity" in IJCA

S.A.Kinariwala working as Assistant Professor in the Computer Science and Engineering Department,MIT Aurangabad. She is Life member of CSI and ISTE .She has published/ presented paper on" Biometric Techniques for Human Identity" in National Conference, Kopargaon, "Information and Network Security" in National conference,NCCITA , PCE,Thane and "Data Mining and Warehousing" in National Conference , Jalgaon , "A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity" in IJCA