Preserving Data Privacy Through Secrecy views in Data Mining

Sailekshmi B
Department of computer Science and Engineering,
Mar Baselios College of Engineering
& Technology
Kerala, India

Ashwini B.
Department of Computer Science and Engineering,
Mar Baselios College of Engineering
Kerala, India

Abstract— **Current inventions in the field of communication technologies and several other technologies such as biometric technologies have given rise to a new research span, known as Privacy Preserving Data Mining (PPDM). It is around for a couple of years now and has accepted on the grounds that it permits exchange of private or confidential data for study purposes. Different algorithms regarding data mining, incorporating these mechanisms, have been developed which allows taking out relevant information from massive amount of data, while hiding sensitive information or data from disclosure or inference. For privacy preservation, different techniques have been proposed such as cryptographic techniques, k-anonymity, data perturbation, anonymization etc. But they suffer from various types of attacks such as linkage attacks, background attacks, homogeneity, integrity loss, information loss etc. Proposed framework is a combination of approaches for privacy conservation in data mining which uses the combined techniques of randomization using matrix of probability and generalization which may reduce the integrity loss and information loss and here pattern of the data cannot be identified by the attacker.**

Keywords— Privacy Preserving Data Mining; k-anonymity; anonymization.

I. Introduction

Nowadays different organizations, companies and industries collect and store very large amount of data for their own needs. These huge amounts of data are then gone for analysis purposes; to obtain relevant or useful information with the help of data mining. After this stage, company or organization will get useful data according to their needs but these types of data or information may include private information or confidential information or sensitive information about individuals.

Privacy can be considered as the right to be abandoned and it is the right to be free from any vigilance and any illogical personal invasions. Couple of cases exists similar to the above condition. For example, statistical data provided by the census bureau collected from houses and individuals are given to third parties for research purposes. If personal details in the statistical data are not hidden, then the third party will easily get information of individual and so privacy breach occurs. Privacy becomes a relevant issue when data set contain confidential details of individuals.

To sort out this problem, PPDM was introduced. For preserving data privacy, a new method was introduced known as PPDM which has been accepted widely [16]. And so, sharing of private data can be permitted for different purpose such as analysis [1]. Main objective is exploring useful data from huge data set and at the same time it provides protection for sensitive information. Two fields of this technique are knowledge hiding and information hiding. Both deals with hiding information, on which knowledge hiding focuses hiding private information and data hiding focuses on modification of sensitive information or removal of private data [3].

There exits several methods for PPDM which can be typically branched into two: cryptographic techniques and non-cryptographic techniques [16]. Cryptographic techniques are the techniques through which data that are sensitive can be encrypted. It is the preferred technique used to yield privacy to the information. Cryptographic technique is very successful approach for the reason that it take care of safety and security to confidential attributes. Other technique i.e. Non-cryptographic techniques involves k-anonymity technique, LKC privacy techniques etc. Here the main problem of non-cryptographic technique is loss of information.

Problems due to privacy and problems due to information loss have a direct relation. To handle this, a new way of preservation should be introduced. Proposed approach can be classified into two segments. In segment 1, randomization (data modification to provide privacy) is applied on original data with the help of matrix of probability. In next segment, the randomized output is then divided into two, based on one private attribute (in the data set as confidential details) as well as non-sensitive details [17]. And then generalization is applied only to confidential attributes to avoid over anonymization (over anonymization results in information loss and integrity loss). By this way, information loss and loss of integrity can be limited to a certain extent. When anonymization is done with the help of probability of matrix and generalization, then it is troublesome for the invader to attack.

Organization of this paper is in this fashion. Segment 2 lay out the literature review. Segment 3 explains the problem definition. Segment 4 covers proposed approach and lastly the segment 5 concludes the work.

II. Literature Review

There exist several methods for PPDM. Goal of this segment is to analyze current approaches in PPDM and to identify drawbacks of those techniques. Existing systems can be classified into cryptographic techniques and non-cryptographic techniques.

Anonymization: Anonymization is an approach for masking confidential data from original or owner's record. Anonymization can be generally classified into generalization and perturbation. Some advantages over cryptographic methods are easy to implement [6]. Some limitations are that they don't guarantee privacy and sensitive data are not preserved properly (because knowledge attacks and homogeneity attacks of k-anonymity algorithm can't uphold confidentiality of information) [6][20].

Cryptographic Technique: Cryptography is an approach through which data that are sensitive can be engrafted. It is the most popular approach used for providing privacy to the information because it provides safety and security to sensitive attributes [1][20]. Some disadvantages are it fails to protect the output data while computation takes place. This method does not give beneficial results when the data set is very large in number (It is a challenging task, to employ this method for vast databases because as instances increases the chances of occurrence of error also increases). Final product may crack the confidentiality of person-specific record because it fails to protect the original data though it is encrypted [1][20].

Some of non-cryptographic techniques are described here such as k-anonymity, data perturbation, l-diversity etc.

A. K- Anonymity

The *k*-anonymity model can be considered as a framework implementing or constructing algorithms and evaluating those systems and algorithms that gives data. The released or publically available data, limits to- what cannot be revealed and what can be revealed about the data entities. For example: to identify a person and the only data available is date of birth and place - there should be at least k number of people meeting with the same requirement [2].

Place	DOB	Gender	Pin code	Disease
Vadassrekonam	1956	M	695143	Stomach Cancer
Puthenthope	1982	M	695586	Myocardial Infraction
Vadassrekonam	1990	M	695143	Stroke
Puthenthope	1982	M	695586	Myocardial Infraction
Vadassrekonam	1956	M	695143	Stomach Cancer
Vadassrekonam	1990	M	695143	Stroke

Table 2.1 An example for k-anonymity

Each data must be in a fashion that any combination of values in each tuples cannot be used for identifying individuals. In this method, the fragments of data representation are reduced by the help of generalization and suppression. Some advantages of this method are maintained data integrity (maintaining the consistency of data) and data granularity (granularity of data refers to the size of data) is reduced (and by that way it prevents the possibility of indirect identification) [2]. Some disadvantages are dealing with large number of quasi identifiers could be problematic and it generalizes or suppresses quasi identifier (set of minimal attributes which can be used to combine additional information to regain the identity of individuals) attributes or demographic attributes (pin code, age, gender) to protect data which reduces quality of data [2]. And there are chances of temporal attacks (due to dynamic collection of data), unsorted matching attacks (due to the ordered arrangement of data) and complementary release attacks [2]

B. Data Perturbation

Data perturbation is a highest accepted approach in PPDM or privacy-preserving data mining. Data Perturbation is an approach for customizing data using random process. These approaches alter confidential data values by modifying them by addition of values or subtraction of values or by using any other mathematical equations [20]. Some advantages of this method are geometric perturbation can conserve the most critical geometric properties [3][18]. Some limitations of this method are they doesn't guarantee privacy for multi-dimensional perturbation (because perturbs multiple columns in one transformation) and other than privacy preservation, accuracy preservation is also considered as a problem [3].

C. L-Diversity Algorithm

The l-diversity model was designed to handle some weaknesses in the k-anonymity model [19]. L-diversity provides privacy even when the publisher of the data does not know what form of information is possessed by the attacker [20]. Some advantages of this method are l-diversity doesn't require information of the entire distribution of the sensitive and non-sensitive attributes (previous method requires the knowledge). L-Diversity does not require the publisher to have as much information as the attacker [5][20]. Some limitations of l-diversity algorithm are, it cannot be used for multiple sensitive attribute, data quality is degraded (when the data is high dimensional).

D. Hybrid Approach

Privacy conservation is a very huge field. Many algorithms such as k-anonymity, data perturbation, l-diversity algorithm etc have been proposed in order to secure the data. Hybrid approach is a new facility through which one can combine two or more approaches to preserve the data [20].

One of the hybrid technique proposed was the combination of randomization and generalization [20]. In this case, data is randomized (with the help of random probabilities) and then generalized the data (modified or randomized data).

Other hybrid techniques proposed was the combination of anonymization and suppression. In this case, data is anonymised and then privacy is given to the modified data.

Some advantages of this technique are the method protects personal data with more preferable accuracy (over anonymization is avoided and due to that reason misrepresentation of data is avoided); also it can rebuild the original data and provide data with no information loss [9][20].

Several other approaches can also be clubbed together to make a hybrid technique such as Data perturbation, Blocking based method, Cryptographic technique, Condensation approach etc.

Some limitations of the existing hybrid approaches are, it is very hard to employ this algorithm for huge databases (because when instances increases the chances of occurrences of error also increases) [9][20].

III. Problem Statement

Though there exist different techniques in privacy preserving data mining, some shortcomings still exists. Some of them are integrity loss, information loss, temporal attacks, over anonymization, identity linkage, attribute linkage, unsorted matching attacks and problems due to contemporary attacks and large data set. For example, if k-anonymity is used, then it minimizes information loss but there are chances of temporal attacks and unsorted matching attacks. To sort out these problems, hybrid approach of different techniques was introduced.

Main motive of this research work is to boost up hybrid approach of randomization and generalization which partially removes the problems of k-anonymity, randomization and l-diversity. Here it increases the utility of data, reduces loss of information and at the same time it provides privacy also. Since it uses random values for randomization stage, output may result to failure is the basic problem here. Proposed approach offers solution to this problem.

IV. Proposed Approach

Proposed approach is a hybrid one which uses the techniques of randomization and generalization. Solution to half of the problems mentioned above (problems due to k-anonymity) can be simplified by using the techniques of randomization with the help of matrix of probability. Data patterns cannot be analysed by the invader if matrix is calculated in randomization.

Randomization is a traditional method for distorting data in privacy preserving data mining. Values in attributes are covered-up here. Existing methods in randomization are additive randomization, multiplicative randomization and micro data randomization. Main idea behind this is to pervert data so that invader cannot determine the data pattern. Proposed approach is a new way of doing randomization using matrix of probability.

Proposed approach can be categorized into two different segments. On segment 1, randomization is performed on original data with the help of probability of matrix. Matrix of probability is the probability of occurrence of each instance under conditions. And in next segment, generalization is performed on randomized output i.e. the output of the section 1 is given to the next segment as input. Generalization is a process of anonymization to hide sensitive values. Flowchart of the proposed solution is given in Fig 4.1.

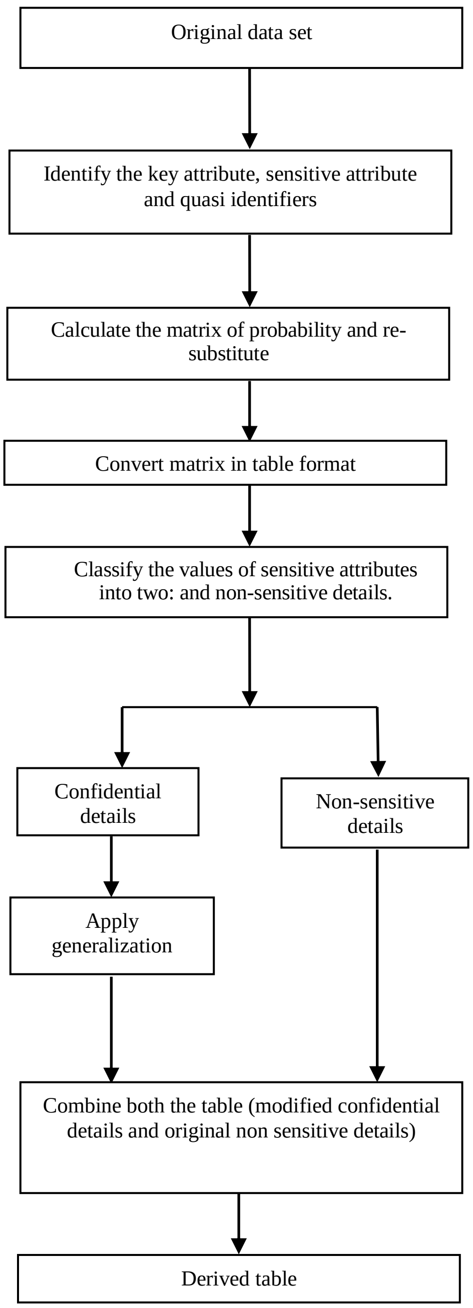
Original data set
Identify the key attribute, sensitive attribute and quasi identifiers
Calculate the matrix of probability and re-substitute
Convert matrix in table format
Classify the values of sensitive attributes into two: and non-sensitive details.
Confidential details
Non-sensitive details
Apply generalization
Combine both the table (modified confidential details and original non sensitive details)
Derived table

Fig 4.1: Flowchart of proposed approach

A. *Algorithm for segment 1*

Input: Original data set

Output: Modified table

Method:

step 1. Identify the key attribute, sensitive attribute and quasi identifiers from the original table.

step 2. Calculate the matrix of probability

a. Select quasi identifiers and sensitive attributes. (Let the number of quasi identifiers be n (n=n1, n2... nn) and the number of records in the original table be j. There will be n number of matrix each of size j*j).

b. Calculate the probability of occurrence of each instance under conditions in a matrix format.

c. Re-substitute the values.

step 3. Convert matrix in table format.

In segment 1, after selecting quasi identifiers, confidential attributes and key attributes; the matrix of probability should be calculated. Matrix of probability is the probability of occurrence of each instance under conditions. For example, consider the following table for calculating matrix of probability. This table is just for an example which cannot be considered as an original table.

Quasi Identifier					Sensitive Attribute
Age	Gender	Symptom 1	Symptom 2	Symptom 3	Disease
33	M	Cough	Chest pain	Breathing Trouble	Lung Cancer
29	F	Cough	Weight loss	Coughing out blood	TB
21	M	Cough	Fatigue	Breathing Trouble	Bronchit is
31	M	upper abdominal pain	indigestio n	vomiting	Chronic Gastritis
60	M	Cough	Chest pain	Breathing Trouble	Lung Cancer
25	F	upper abdominal pain	indigestio n	vomiting	Gastritis

Table 4.1 Medical data

There are 5 quasi identifiers: age, gender, symptom1, symptom2 and symptom 3. Here aim is to calculate the probability of occurrence of each instance. For that each instance should be considered.

Consider the 1st attribute age, the 1st quasi identifier. Each value for age i.e. 33, 29, 21, 31, 60, 25 should be considered with every row i.e. to calculate the probability of occurrence of each age under conditions such when the gender is “so and so”, when the symptom 1 is “so and so”, when the symptom 2 is “so and so”, when the symptom 3 is “so and so”. The probability of occurrence of age 33 under conditions such when the gender is “M”, when the symptom 1 is “Cough”, when the symptom 2 is “Chest pain”, when the symptom 3 is “Breathing Trouble” should be calculated first. After that next row should be considered at this same age. The probability of occurrence of age 33 under conditions such when the gender is “F”, when the symptom 1 is “Cough”, when the symptom 2 is “Weight loss”, when the symptom 3 is “Coughing out blood” should be calculated next. Similarly calculate every row till last one. The probability of occurrence of age 33 under conditions such when the gender is “F”, when the symptom 1 is “upper abdominal pain”, when the symptom 2 is “indigestion”, when the symptom 3 is “vomiting” should be calculated.

Now 2nd value for age should be considered for every row. Calculate the probability of occurrence of age 29 under conditions such when the gender is “M”, when the symptom 1 is “Cough”, when the symptom 2 is “Chest pain”, when the symptom 3 is “Breathing Trouble”. After that next row should be considered at this same age. The probability of occurrence of age 29 under conditions such when the gender is “F”, when the symptom 1 is “Cough”, when the symptom 2 is “Weight loss”, when the symptom 3 is “Coughing out blood” should be calculated next. Similarly calculate every row till last one. The probability of occurrence of age 29 under conditions such when the gender is “F”, when the symptom 1 is “upper abdominal pain”, when the symptom 2 is “indigestion”, when the symptom 3 is “vomiting” should be calculated.

Similarly each value for age should be calculated for every row till occurrence of age 25 under conditions such when the gender is “F”, when the symptom 1 is “upper abdominal pain”, when the symptom 2 is “indigestion”, when the symptom 3 is “vomiting” should be calculated. This is the procedure for calculating probability of matrix for 1st quasi identifier. Similarly each instance should be considered for every quasi identifier.

Calculation of probability of occurrence can be explained with an example. Consider the condition “The probability of occurrence of age 33 under conditions such when the gender is “M”, when the symptom 1 is “Cough”, when the symptom 2 is “Chest pain”, when the symptom 3 is “Breathing Trouble” ”.

Example for matrix of probability: -matrix 1

0.3333333333333333333333333333	0.3333333333333333333333333333	0.2666666666666666666666666667	0.1333333333333333333333333333	0.2666666666666666666666666667	0.0666666666666666666666666667
0.2666666666666666666666666667	0.0666666666666666666666666667	0.0666666666666666666666666667	0.2666666666666666666666666667	0.3333333333333333333333333333	0.3333333333333333333333333333
0.3333333333333333333333333333	0.1333333333333333333333333333	0.1333333333333333333333333333	0.1333333333333333333333333333	0.0666666666666666666666666667	0.0666666666666666666666666667
0.0666666666666666666666666667	0.0666666666666666666666666667	0.1333333333333333333333333333	0.0666666666666666666666666667	0.0666666666666666666666666667	0.0666666666666666666666666667
0.2	0.0666666666666666666666666667	0.1333333333333333333333333333	0.2	0.0666666666666666666666666667	0.0666666666666666666666666667
0.1333333333333333333333333333	0.1333333333333333333333333333	0.0666666666666666666666666667	0.0666666666666666666666666667	0.0666666666666666666666666667	0.1333333333333333333333333333

Re-substitution

step 1. To re-substitute the values for each quasi identifier N, consider the corresponding matrix.

step 2. Consider only the diagonal values for every matrix

a. Let N11 be the 1st value of matrix

N11 =p(a)*p(b)*p(c)..

For n1 consider p(a)

step 3. Compare it with mapping table.

a. If value p(a) exists in the table

- Check the duplication of that value.

i. If it exists only 1 time, then just give that answer.

ii. If not, consider all the answers of duplication

○ Num-> generalize

○ String-> use * instead

Example: Let the matrix obtained from the original table selected be the following one

N11	N12	N13	N14	N15	N16
N21	N22	N23	N24	N25	N26
N31	N32	N33	N34	N35	N36
N41	N42	N43	N44	N45	N46
N51	N52	N53	N54	N55	N56
N61	N62	N63	N64	N65	N66

Consider the diagonal values only, because the value according to the table lies diagonally in the matrix. Now consider each diagonal value. From matrix calculation, N11 is the product of n1, n2 … n6 which are actually the probability values. For example consider the values for x1 to x6. It is clear that x1 denotes the exact age.

N11 = n1 * n2 * n3 * n4 * n5 * n6

N11= 1/6 * 4/6 * 4/6 * 2/6 * 3/6 * 2/6

p(1) = 1/6 → age

Similarly the values for x22 to x66 should be calculated.

Consider an example:

Age	Gender	Disease
21	M	HIV+
35	F	Cancer
35	M	HIV+

Let matrix created for age be

j	k	l
m	n	o
p	r	q

Age and probability of occurrence:

Age: 21 →1/3: probability

Age: 35→2/3: probability

Age= p(j) * p(n) * p(q)

=p(occurrence of age 21 under conditions when gender='M' and Disease='HIV+') * p(occurrence of age 35 under conditions when gender='F' and Disease='Cancer') *

p(occurrence of age 35 under conditions when gender='M' and Disease='HIV+')

p(j) = 1\3 * 2\3 * 2\3 → 21

p(n) = 2\3 * 1\3 * 1\3 →35

p(q) = 2\3 * 2\3 * 2\3 →35

Similarly calculate Gender. And then generalize

– Age→ 2\13 - - - > 54, 65, 21, 30 -- → 20-60

– Age→ 4\13 - - - > 73, 87, 90-- → 70-90

B. Algorithm for next segment

Input: Converted table.

Output: Derived table.

Method:

step 1. Select converted table.

step 2. Classify the values of sensitive attributes into two: confidential details and non-sensitive details.

step 3. Consider the confidential details and apply generalization to the details.

step 4. Combine both the table (modified confidential details and original non sensitive details).

step 5.

After completing segment 1, the output of segment 1 is given to next segment as input. In traditional methods of PPDM, generalization is applied to entire data which may result in information loss due to over anonymization. So here in the work, it considers only highly sensitive information and applies generalization only to this section. Rest of the data set is not generalized and kept it as such because in this area, anonymization is not required. If entire data set is generalized, then this second part of data (non-confidential data) is also generalized which is of no use. To avoid this problem confidential attribute can be split into two; confidential details and non-confidential details [21]. Generalization is applied only to confidential details so that anonymization occurs only to those parts. Generalization is actually a process of anonymization which duplicates the record. Here in this work, generalization is done separately for strings and numerical. For strings, each word is modified by changing the second alternative positions (letters) in each word with asterisk values only after finding the total strength of each word. For numerical, last three values are hidden by replacing the original value with asterisk only after calculating the total numbers in each value.

V. RESULT AND DISCUSSION

Previous method of hybrid approach of randomization and generalization had a major drawback which leads to different answers for the same data set. This is because of the probability matrix (here the matrix has been created with random values which are not fixed). Due to this random generation of values, each time when a data set is checked, it will result different values for the same data set. Main advantages are loss of information is reduced, data utility has been increased and at the same time it was able to provide privacy. But the problem is integrity loss. If integrity and accuracy of the data is not been protected correctly, then data will be of no use. The below figures shows the results of same data set when data has been checked for two times. This happens because of the random value generation in the matrix of probability.

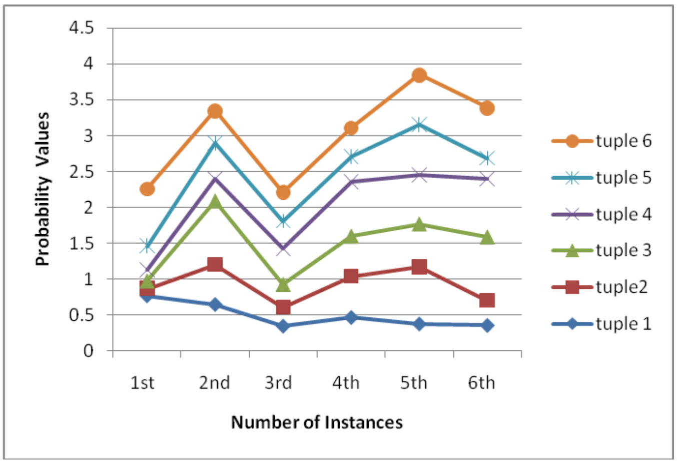

Fig 5.1 Random values of 1st output

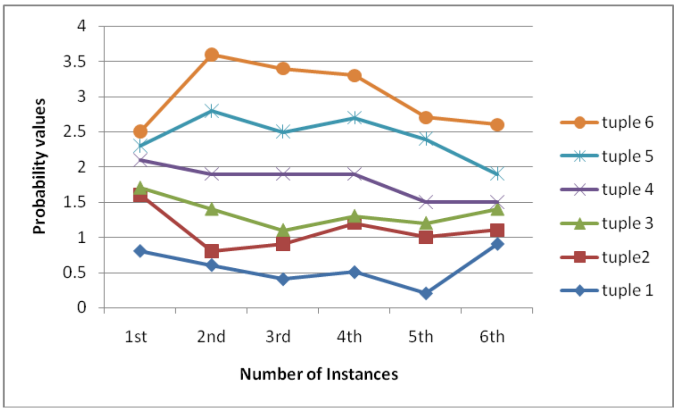

Fig 5.2 Random values of 2st output

From the graph itself it is clear that there is a drastic change in each calculation which finally leads to incorrect result.

This problem has been sorted out by the proposed approach. Here the values of matrix are calculated with the help of probability matrix. The random generation process is omitted here and instead of that original values are calculated. And here the values won't be changed for the same data set.

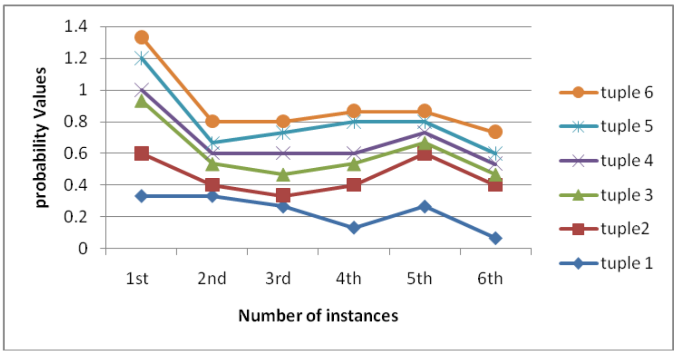

Fig 5.3 Output of Matrix of Probability for 1st and 2nd checking (for same data set)

Result of matrix values are then re-substituted to get a converted table. After re-substitution, the converted table which has been modified is given to section 2 as input. And confidential and non-sensitive data are separated here. Then generalization is only applied to confidential data. The matrix obtained here is given below

```
pmatrix$ 0.3333333333333333333333333333 0.3333333333333333333333333333 0.0666666666666666666666666667 0.2666666666666666666666666667 0.133333333333333333333333333
6666666666666666667 0.0666666666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333
0.1333333333333333333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2 0.2 0.06666666666666666666666
3333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.133333333333333333
.1333333333333333333333333333 0.2 0.1333333333333333333333333333 0.1333333333333333333333333333 0.2666666666666666666666666667 0.1333333333333333333333333333 0.1
3333333333333 0.3333333333333333333333333333 0.0666666666666666666666666667 0.2666666666666666666666666667 0.1333333333333333333333333333 0.266666666666666666666
6666666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.133333333
3333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2 0.2 0.0666666666666666666666666667 0.06666666666666666666
6666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.1333333333333333333333333333 0.0666666666666
333 0.2 0.1333333333333333333333333333 0.1333333333333333333333333333 0.2666666666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333
33333333333333333 0.0666666666666666666666666667 0.2666666666666666666666666667 0.1333333333333333333333333333 0.2666666666666666666666666667 0.066666666666666
.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.13333
66666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2 0.2 0.0666666666666666666666666667 0.0666666666666666666666666667 0.26666666666666
666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.1333333333333333333333333333 0.0666666666666666666666666667 0.066666666
33333333333 0.1333333333333333333333333333 0.2666666666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333 0.333333333333333333333
66666666666666666666667 0.2666666666666666666666666667 0.1333333333333333333333333333 0.2666666666666666666666666667 0.0666666666666666666666666667 0.2666666666
333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1
66666666666667 0.0666666666666666666666666667 0.2 0.2 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.2666666666
.0666666666666666666666666667 0.0666666666666666666666666667 0.1333333333333333333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666
33333333333333333 0.2666666666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333
2666666666666666666666666667 0.1333333333333333333333333333 0.2666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.06666
33333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333
666666666666666666667 0.2 0.2 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.266666
667 0.0666666666666666666666666667 0.1333333333333333333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.1
666666666666666666667 0.1333333333333333333333333333 0.1333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333333333333333 0.33333333333
67 0.1333333333333333333333333333 0.2666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.0666666666666666666666666667 0.0
3333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.133333333333333333
2 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.26
66666667 0.1333333333333333333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2 0.066666666666666666666
3333333333333333333333333 0.1333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333333333333333 0.133333
333333 0.2666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.0666666666666666666666666667 0.06666666666666666666666666
333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.2 0.13333333333
666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667
3333333333333 0.0666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2 0.0666666666666666666666666667 0.13333333333333333
3 0.1333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333333333333333 0.3333333333333333333333333333 0.1333333333333333333333333333 0.33
66666666667 0.0666666666666666666666666667 0.2666666666666666666666666667 0.0666666666666666666666666667 0.0666666666666666666666666667 0.2666666666666666666
333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.1333333333333333333333333333 0.2 0.1333333333333333333333333333 0.13333
6666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.2666666666666666666666666667 0.133333333333333333333
```

CONCLUSION

Privacy, confidentiality and security are primary concern when data is considered. Nowadays society immensely worries about their confidential information given to others for various reasons and because of that reason many of them are not ready to disclose information which may result in false data set. Though there exist different techniques in privacy preserving data mining, some shortcomings still exists. Some of them are integrity loss, information loss, temporal attacks, over anonymization, identity linkage, attribute linkage, unsorted matching attacks and problems due to contemporary attacks and large data set. For example, if k-anonymity is used, then it minimizes information loss but there are chances of temporal attacks and unsorted matching attacks. To sort out these problems, hybrid approach of different techniques was introduced. Main motive of this research work is to boost up a hybrid approach of randomization and generalization which partially removes the problems of k-anonymity, randomization and l-diversity. Here it increases the utility of data, reduces loss of information and at the same time it provides privacy also. But the problem is; since it uses random values for randomization stage, output may result to failure. Proposed approach offers solution to this problem

In future instead of generalization in the second segment, k-anonymity, l-diversity and lkc privacy can be applied to increase the data utility without replacing the first section.

REFERENCES

[1] Y. Lindell and B. Pinkas, Privacy Preserving Data Mining,Journal of Cryptology, Vol. 15, No. 3, pp. 177-206, 2002.

[2] L.Sweeny, “k-anonymity:a model for protecting privacy” International Journal on Uncertanity,Fuzziness and knowledge-based systems, pp.557-570, 2002.

[3] Chen,K. And Liu, “Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining”, Proceedings of International Conference on Data Mining (ICDM), IEEE, 2010.

[4] Z. Zhang and A. Mendelzon, “Authorization Views and Conditional Query Containment”, International Conference on Database Theory (ICDT'05), pp. 259-273, 2007.

[5] Ashwin Machanavajjhala, Johannes, Gehrke Daniel Kifer"ℓ-Diversity: Privacy Beyond k-Anonymity", ACM International Conference on Management of Data (SIGMOD), pp. 551-562, 2011.

[6] S.Shaik Parveen, Dr.C.Kavitha, “Review on Anonymization”, International Journal of Computers & Technology, Volume 3 No. 3, Nov-Dec 2012.

[7] J. Liu, J. Luo and J. Z. Huang, “Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements”, in proceedings of 11th IEEE International Conference on Data Mining Workshops, 2011.

[8] T. Jahan, G.Narsimha and C.V Guru Rao, “Data Perturbation and Features Selection in Preserving Privacy” in proceedings of Conference on Privacy Management of Data, 2012.

[9] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, in proceedings of the Third IEEE International Conference on Data Mining, 2003.

[10] Manish Sharma, Atul Chaudray, Manish Mathuria,Santhosh Kumar, “An Efficient approach for privacy preserving in data mining”, International Conference on Signal Propagations and computer technology(ICSPCT), 2014.

[11] Agarwall, R. and Shrikant, R. “Privacy Preserving Data Mining”, Proceeding of Special Interest Group on Management of Data, pp.439-450, 2000.

[12] Jian Wang, Yong Cheng Luo, Yen Zha, Jiajin Le, “A Survey on Privacy Preserving Data Mining”, International Workshop on Database Technology and Applicationpp.111-114,2009.

[13] V.S Verkoys, A.K Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, “Assosiation Rule Hiding”, IEE Transaction Knowledge and Data Engineering, 16(4); 434-447,2004

[14] P. Samurai, L Sweeny, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression”, Technical Report, SRI International,1998.

[15] A.Agarwal and R. Srikant, “Privacy Preserving Data Mining”, ACM SIGMOD International Conference on Kmowledge Discovery and Data Mining, vol.29 no.2,pp. 50-57,2004

[16] Fung, Benjamin C. M.; Ke Wang; Rui Chen and Yu, Philip S.. "Privacy-Preserving Data Publishing: A Survey of Recent Developments" , ACM Computing Surveys, 2010.

[17] Li Xiao-Bai Sarkar, Sumit. "Privacy protection in data mining: a perturbation approach for categorical data." , Information Systems Research, Sept 2006 Issue

[18] Keke Chen. "Geometric data perturbation for privacy preserving outsourced data mining" , Knowledge and Information Systems, 2010.

[19] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam.” -Diversity: Privacy Beyond k-Anonymity” ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, March 2007.

[20] www.cse.psu.edu and www.ijcsit.com

[21] Savita Lohiya, Lata Ragha, “Performance Analysis of Hybrid Approach for PrivacyPreserving in Data Mining”, Int. J. on Recent Trends in Engineering and Technology, Vol. 8, No. 1, Jan 2013

[22] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.