International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 10, ISSN No. 2455-2143, Pages 310-316
Published Online February 2020 in IJEAST (http://www.ijeast.com)

A REVIEW ON ANONYMIZATION TECHNIQUES FOR PRIVACY PROTECTION IN DATA MINING

U.Saranya,
Department of MCA,
Dr. Mahalingam College of
Engineering and Technology,
Tamilnadu, India.

A.Logeswari,
Department of IT,
Dr. Mahalingam College of
Engineering and Technology,
Tamilnadu, India.

U.Sujatha,
Department of MCA,
Dr. Mahalingam College of
Engineering and Technology,
Tamilnadu, India.

Abstract - Data anonymization is a primary technique used purposely for the protection of privacy. To remain anonymous, it excludes personal identifiable information from data sets. Preservation of privacy is a major factor to be considered in protecting against attacks by unauthorized entities on identity disclosure and ensuring that data is anonymized and still efficient for analytical tasks. The data can be released to ensure that the loss of information is minimal in order to maintain the utility of the analysis for further tasks. The data in the dataset may be a mixture of sensitive and non-sensitive information. To order to protect sensitive data from the outside world, privacy protection strategies for data publishing are rapidly increased. K-anonymity is one of important privacy protection strategies, but more attention needs to be paid to increasing data usefulness and more loss of knowledge before publishing. Using L-diversity and T-closeness strategies, privacy security can be based on nodes. Edge perturbation and edge randomization are privacy conservation techniques in social network graphs. Relational information and social network data protect privacy using K-anonymity, edge perturbation, edge randomization, and L-diversity techniques. In this paper, a comparative review and study on K-Anonymity, L-Diversity, T-Closeness and perturbation Anonymization techniques is presented along with slicing for high-dimensional databases and procedure for following the reduction of dimensionality with selection of features.

Keywords: **K-anonymity, L-diversity, Privacy preserving data publishing, T-closeness**

I. INTRODUCTION

Anonymization of information plays a major role in protecting data sets confidentiality before data is released. The published data must not permit unauthorized users to know anything about the target persons. This paper provides a survey and study of various privacy protection anonymization methods such as k-anonymity, l-diversity, t-closeness. The k-anonymity issue based on expanded data set with reduced utility. Attacks like Homogeneity Attack and Background Knowledge Attack will occur in K-anonymity. The L -diversity can be applied to protect the data with an increase in the data set. L -diversity eliminates information losses in data sets of K-anonymity as data point moves any volume. The privacy protection strategy T –closeness allows that the distribution of a critical attribute in any equivalence group be identical to the distribution of the attribute in the overall table to avoid disclosure of attributes.

Healthcare, retail, digital media, financial, and beyond companies rely on data that includes personally identifiable information (PII). The identifiable information, such as age, gender, name, etc., is changed or removed from a set of data to prevent the determination of the individual to which the data belongs. The term is commonly referred to as "data sanitization" or "data masking."

Anonymization of data is one of the effective ways to enhance data sharing integrity Anonymization of information reduces the risk of exposure by sharing data between countries and industries worldwide. For example, sharing confidential data about a hospital on its patients to a medical research laboratory or pharmaceutical company needs to be hidden keeps patient's anonymous information by removing names, social security numbers, birth dates and addresses of their patients to provide the components needed for medical research such as age, ailments, height, weight, sex, race, etc. The publisher should check the privacy of anonymity before publishing the data.

II. LITERATURE REVIEW

A. Dataset Description

The performance of the proposed algorithm based on Chaos and Perturbation Techniques [1] is assessed on the data set for adults extracted from the U.S. 1994. Census database This data set is used in this study because it is used as a benchmark for the analysis of algorithms in the literature. The data set is available online from the University of California-Irvine

Machine Learning Repository. This includes 32,561 records and the total number of unmissed records is 30,162. The number of attributes is 15. In the data set, 7508 instances are in class ">50 K" and 22,654 instances are in class "≤50 K".

To illustrate the scalability of the proposed Big Data algorithm, the Adult Data Set is expanded consistently with records of ~60 K, 120 K, 240 K and 480 K respectively. Information replication is conducted to determine the reliability of the classification without corrupting data integrity.

B. Survey on data anonymization Techniques for large data sets

The data stored in the cloud can contain information unique to the user [2]. To maintain the privacy of the client, this data must be secured. The information includes clear identification, sensitive identification, non-sensitive identification and quasi identification. The clear identifier provides direct information about the record holder and the combination of quasi-identifier distinguishes the user's specific data from the dataset. The important identification contains information of income, health issues, etc. to record holders. The non-sensitive identifier belongs to all the other set of data.

Anonymization technique is used to cover these sensitive data as long as it is appropriate to preserve these user-specific data for future analysis [3]. Attacks such as linking attacks can be avoided by anonymizing the quasi identifier before the information is released. The quasi-identifier is updated using QID (quasi-identifier) attributes in the original table. If the data holder is still connected using the updated QID, then several records are mapped to make the relation unclear. There are many other types of attacks on the data set that can be handled using the correct techniques for anonymization.

III. ANONYMIZATION TECHNIQUES

A. K-anonymity

K-anonymity is a model of privacy conservation in which each record published on its QI attribute must be indistinguishable from at least (k-1) others. The "quasi-identifiers" are the attributes available to an adversary such that a table T satisfies k-anonymity if there are k−1 other tuples ti1, ti2,. To the degree that t[C] = ti1 [C] = ti2 [C] =. For all C = tik−1 [C].[3]

The protection provided by k-anonymity techniques is simple, and if a table satisfies k-anonymity for some value k, then anyone who knows only one individual's quasi-identifier values can not identify that individual's corresponding record with confidence greater than 1/k[4]. While k-anonymity protects from disclosure of identity, it does not provide adequate protection against disclosure of attributes. Many researchers have noted this, e.g. [5, 8, 11]. Two attacks have been identified:

Homogeneity attack
Background Knowledge attack.

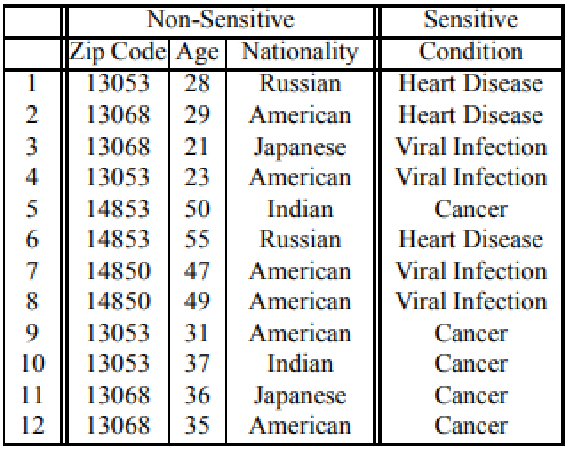
Homogeneity Attack

Alice and Bob are neighbors that are antagonistic. One day, Bob gets ill and is taken to the hospital by ambulance. Having seen the ambulance, Alice is trying to find what disease Bob is suffering from cancer. Alice finds the hospital's 2-anonymous list of current hospital information (Table 2), and she knows that one of the records in this table includes Bob's data [5]. Since Alice is a neighbor of Bob, she knows that Bob is a 31-year-old American male living under zip code 13053. Alice therefore assumes that the record number of Bob is 5, 9, 10, 11 or 12. All these patients now have the same medical condition (cancer), and Alice assumes that Bob has cancer. This is the homogeneity attack.

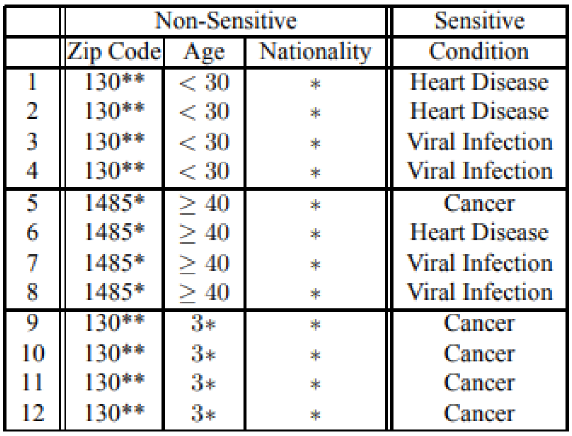
Observation 1

K – Anonymity can create groups that leak information in the sensitive attributes due to lack of diversity. Suppose we have a dataset with 60,000 separate tables as a back of the envelope calculation where the sensitive attribute will take 3 distinct values and is not associated with any sensitive attributes. A 5-anonymization of this table will have about 12,000 groups 2 and 1 out of every 81 groups will have no diversity on average. Around 148 groups with no diversity should expect the result. Data around 740 individuals would therefore be affected by an assault on homogeneity.

Table 1. Original data



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

The sanitized table should also ensure "diversity" in addition to k-anonymity, all tables that share the same values of their quasi-identifiers should have different values for their sensitive attributes [5]. The next point is that an opponent could use knowledge of "Background" to uncover sensitive information.

Background Knowledge Attack

For an example of the background knowledge attack, assume that Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 2 by knowing Carl's age and zip code. Suppose Alice knows Carl's risk of heart disease is very small. This knowledge makes it possible for Alice to believe that Carl is most likely to have cancer.

Alice has a pen friend named Umeko who has been admitted to the same hospital as Bob and whose patient records are also shown in table 2. Alice knows that Umeko is a Japanese woman who is 21 years old and currently lives in code 13068. Alice finds from this data that Umeko information is in record number 1, 2, 3, or 4.

Alice is not sure if Umeko caught a virus or has HIV without additional information. Japanese people are well known to have an extremely low rate of cardiac disease. Therefore, Alice assumes that Umeko has a viral infection with almost certainty.

Observation 2

K – Anonymity does not guard against Background knowledge-based attacks. This has shown that a k-anonymous table can disclose sensitive information. Since both of these attacks are appropriate in real life, a stronger concept of privacy is required, taking into account diversity and knowledge of background [3]. The definition of l-diversity was developed in order to avoid the above problems of k-anonymity attacks.

B.**L-diversity**

L-diversity is a group-based model of anonymization that helps to preserve data privacy by reducing data representation granularity by generalizing and suppressing data [6]. In l-diversity, a class of equivalence is said to have l-diversity if the sensitive attribute has at least l "well-represented" value. A table is said to have l-diversity if each table equivalence class has l-diversity and if it contains at least l "well-represented" values for the sensitive attribute S, a q block is l-diversity. If each q block is at least l-diversity [6, 3], a table is l-diversity.

The disadvantage of k-anonymization due to the knowledge attack can be eliminated by diversifying the sensitive attribute values within a block. The l -diversity template prevents privacy protection attribute disclosure [2].

Properties

It is not necessary to know the full distribution of the sensitive and non-sensitive attributes in l-diversity.

L-Diversity does not even allow the data publisher to have the same details as the opponent. The higher the value of l, the more information is needed to exclude potentially sensitive attribute values.

Various opponents may have different background information which leads to different inferences. This defends against all of them simultaneously without the need for inferences that can be made with Background Knowledge [5].

Distinct L-diversity

In the concept of l -diversity, the word "well represented" reflects that in each equivalence class there are at least l distinct values for the sensitive attribute. Distinct ℓ -diversity does not prevent probabilistic inference attacks. In such a case, one value may appear much more frequently in an anonymized block than other values, allowing an adversary to conclude that it is very likely that an entity in the equivalence class will have that value.

Entropy L-diversity

Equivalence class E entropy is defined

$E = -p (E, s) \log p (E, s)$

Where S is the sensitive attribute's domain, and where p (E, s) is the fraction of records in E with sensitive value s. If for each equivalence class E, Entropy (E) is log l, a table is said to have entropy l -diversity. Entropy l-Diversity is stronger than l –diversity. The entropy of the entire table must be at least log (l) to have entropy l -diversity for each equivalence group. The responsive QI attributes will differ from group to group, so the semantic closeness of these values is not taken into account[6]. If a few values are very common, the entropy of the entire table may be small.

Recursive (c, ℓ)-diversity

Recursive (c, l)-diversity means that the most common value does not appear too often and that the less frequent values do not appear too uncommon. Let m be the number of values in a class of equivalence, and ri, 1 ≤ i ≤ m be the number of times that the ith most frequent sensitive value appears in an equivalence class E. Then E is said to have recursive (c, ℓ)-diversity if r1 < c (rl +rl+1 +...+rm). A table is said to have recursive (c, l)-diversity if it has recursive (c, l)-diversity in all its equivalence classes [7].

Limitations of L-Diversity

There are certain limitations to the step beyond k-anonymity to protect against disclosure of attributes. L-diversity can be complicated and needless.

Suppose the original data has one important attribute only: the test result for a specific virus. Two values are needed: positive and negative. Suppose there are 10,000 records, 99% of which are negative and only 1% positive. Then the two values have different resistance points. One wouldn't mind being assessed negatively, because then another is the same as 99% of the population, but one wouldn't want to be known / considered positive. In this case, for an equivalence class that contains only records that are negative, 2-diversity is unnecessary[10]. There may be a limit of 10000×1% = 100 equivalence groups to have a distinct 2-diverse table, and the knowledge loss would be high.

The entropy of the responsive attribute in the overall table is very high; the value must be set to a small. L-diversity is inadequate to prevent disclosure of attributes. Two attacks listed below are the main concern about L-diversity.

Skewness Attack

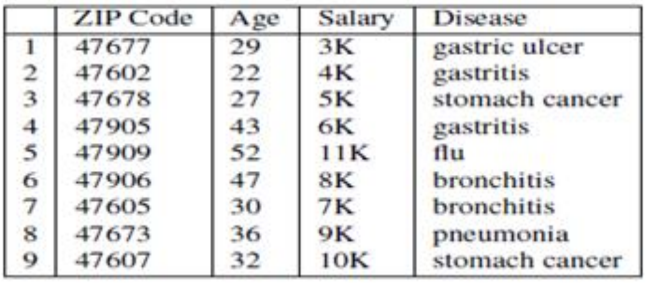
Satisfying L-diversity does not preclude disclosure of attributes when the total distribution is distorted. Remember the example of the virus again. Suppose one equivalence group has the same number of positive records and negative records. It fulfills separate 2-diversity, 2-diversity entropy and any recursive (c, 2)-diversity condition that can be imposed. The risk of confidentiality may be increased because anyone in the class would be perceived to have a 50% chance of being optimistic relative to 1% of the overall population.

Find now a class of equivalence with 49 positive records and only 1 negative record. It would be noticeably 2-dimensional and entropy higher than the overall table, but anyone in the equivalence group would be considered optimistic at 98%, rather than 1%. In addition, this equivalence class has almost the same composition as a class with 1 positive and 49 negative record, although there are very different levels of privacy risks in both categories.

Similarity Attack

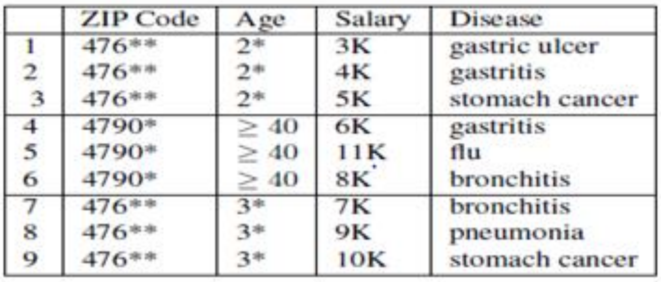
If the sensitive attribute values are distinct but semantically identical in an equivalence group, an opponent will learn important information. Consider the example below. Example 3 Table 3 is the original table, and Table 4 displays an anonymous version that satisfies the distinct and entropical 3-variety. There are two characteristics that are sensitive: salary and disease. Suppose one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's wage is within [3K–5K] range and can deduce that Bob's wage is relatively low. This attack not only applies to numerical attributes such as "Salary," but also to categorical attributes such as "Disease". Assuming that Bob's record belongs to the first equivalence group, one can infer that Bob has some stomach-related problems because all three diseases in the class are related to the stomach.

Table 3. Original/Salary Table



	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 4. A 3-diverse version of Table 3



	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

This leakage of sensitive information occurs because while the criterion for L-diversity guarantees "diversity" in each class of sensitive values, it does not take into account the semantic closeness of these values [8].

Distributions with the same level of diversity can provide very different levels of privacy, as there are semantic associations between the values of attributes, as different values have very different levels of sensitivity, and therefore privacy is also influenced by the relationship with the overall distribution.

C. Suppression

Input: Compliance with and non-compliance with node.

Output: Instance in anonymized data set.

Step1: Calculate how many instances for each complying node; it can account for non-complying nodes.

Step2: If the number of instances which the complying node can compensate is greater or equal to the number of required instances then compensation is possible, perturbation will be performed.

Step3: In compensation the number of instances needed from the complying node to the non-complying node. Non-complying node instances are then transferred to anonymized dataset with the quasi-attribute values suppressed and the remaining instances of the complying node are shifted to the anonymized dataset [8].

D.T-Closeness

A new measure of confidentiality and security is measured by an observer's benefit of data. The observer has an idea of the sensitive attribute value of a person before the data is published. The analyst has a later conviction after seeing the table published. Knowledge benefit can be interpreted as the difference between the belief afterwards and the belief beforehand. The approach is focused on separating the gain of information into two parts: that of the entire population in the data released and that of specific individuals.

Perform the following experiment to activate the technique: First, an observer has some prior conviction B0 about the sensitive attribute of a person. Then the observer is given a fully generalized version of the data table in a hypothetical step where all attributes in a quasi-identifier are removed. The view of the observer is determined by Q, the distribution of the sensitive value of the attribute throughout the table, and changes to B1[8]. Finally, the table released is given to the observer. By understanding the individual's quasi-identifier values, the observer may define the equivalence class in which the records of the individual are in, and learn the P distribution of sensitive attribute values in this class [12].

The observer's belief changes to B2. The criterion for L-diversity is based on the constraint of the gap between B0 and B2Choose to reduce the B1-B2 gap. This interprets Q, the distribution of the critical attribute in the table's overall population, as public information. It does not restrict the information gained by the observer about the population as a whole, but it restricts the degree to which the observer can obtain more information about specific individuals. To support the statement that Q should be regarded as public information, it is observed that generalizations and all quasi-identifiers have the most general meaning attributes. Until the launch of a version of the data, a Q distribution will be released. It also claims that if you want to release the table at all, you intend to release the Q distribution, which makes information useful in this table. In other words, Q is intended to be public information. A major change from B0 to B1 means a lot of new information is contained in the data table, e.g. the new data table corrects some widely held misconceptions. If the gap between B0 and B1 is larger, the more useful the information will be collected. Since the gain of information between B0 and B1 affects the entire population, do not restrict this profit. By restricting the range between P and Q, the gain can be restricted from B1 to B2. When P = Q is the same, then B1 and B2 are the same. If P and Q are similar, B1 and B2 should also be close, even though B0 could be very different from both B1 and B2.

The definition of t-closeness is given as an equivalence class if the difference between the distribution of a sensitive attribute in this category and the distribution of the attribute throughout the table is only a threshold t. If all equivalence classes have t-closeness, a table is said to have t-closeness.

E. Perturbation

Input: non complying node.

Output: Instance in anonymized data set.

Step1: Find the value of the parent node splitting attribute for each non-complying node.

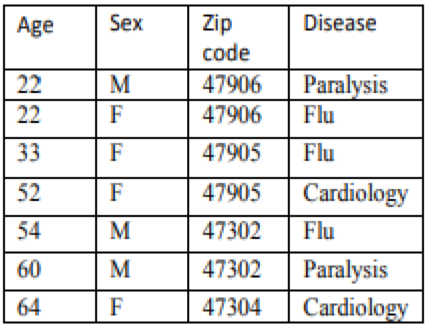
Step2: Perturb by applying half of the non-complying node instance value to the parent node attribute value.

Step3: Shift instances with the quasi attribute values suppressed to the anonymized dataset [9].

F. Slicing

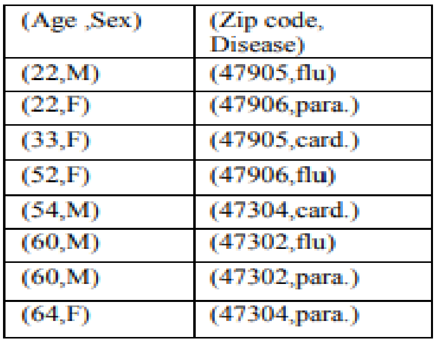
In generalization and bucketization, T., to address problems. Li introduce a new technique of slicing [9] to protect publishing confidentiality. Slicing based on data partitioning in this technique. Partitioning is done both horizontally and vertically. Highly correlated attributes are grouped into columns in vertical partitioning. Every column has a subset of highly correlated attributes. Tuples are organized into seals in horizontal partitioning. olumn values are sorted randomly to break the link between different columns. For example, Table 6 is sliced data of Table 5.

Table 5. Original Data



Age	Sex	Zip code	Disease
22	M	47906	Paralysis
22	F	47906	Flu
33	F	47905	Flu
52	F	47905	Cardiology
54	M	47302	Flu
60	M	47302	Paralysis
64	F	47304	Cardiology

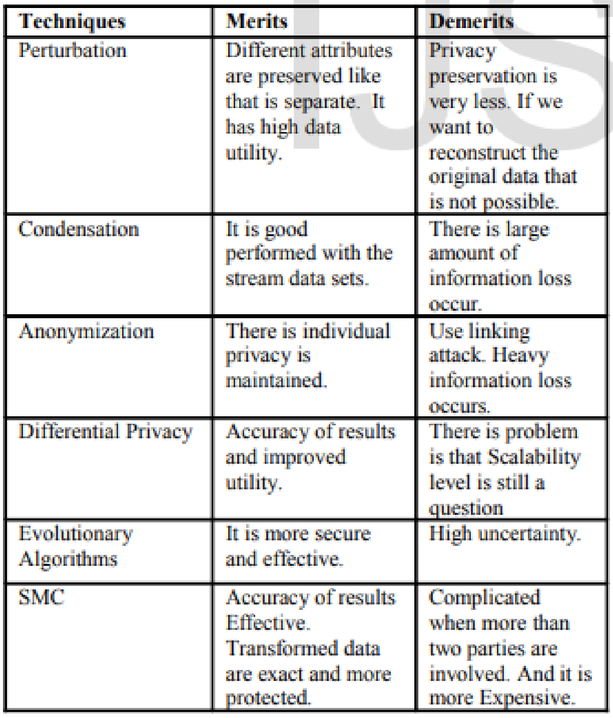
Table 6. A published Data by Slicing



(Age ,Sex)	(Zip code, Disease)
(22,M)	(47905,flu)
(22,F)	(47906,para.)
(33,F)	(47905,card.)
(52,F)	(47906,flu)
(54,M)	(47304,card.)
(60,M)	(47302,flu)
(60,M)	(47302,para.)
(64,F)	(47304,para.)

Slicing reduces the dimensionality of the data. Conserves better usefulness compared to generalization and bucketization. Slicing together not only groups of highly correlated attributes, but also holds the associations between attributes. It breaks the link between uncorrelated attributes and gives more privacy to the publication of data. Because these characteristics are not unique, this is a simple task to classify. Slicing offers better protection of privacy because each tuple has more than one game[6]. Slicing can be used effectively to prevent disclosure of attributes. Slicing retains better data quality compared to generalization. Slicing can also handle high-dimensional data.

Table 7. Merits and demerits of Anonymization techniques



Techniques	Merits	Demerits
Perturbation	Different attributes are preserved like that is separate. It has high data utility.	Privacy preservation is very less. If we want to reconstruct the original data that is not possible.
Condensation	It is good performed with the stream data sets.	There is large amount of information loss occur.
Anonymization	There is individual privacy is maintained.	Use linking attack. Heavy information loss occurs.
Differential Privacy	Accuracy of results and improved utility.	There is problem is that Scalability level is still a question
Evolutionary Algorithms	It is more secure and effective.	High uncertainty.
SMC	Accuracy of results Effective. Transformed data are exact and more protected.	Complicated when more than two parties are involved. And it is more Expensive.

Advantage of slicing

Slicing ensures better data quality compared to generalization

Slicing is more efficient than bucketizing.

Slicing can also manage data of large dimensions.

IV. DIMENSIONALITY REDUCTION USING FEATURE SELECTION

The reduction of dimensionality using feature selection algorithm is used to efficiently handle high-dimensional data. Pre-processing is carried out on the high-dimensional data set to handle missing values. Typically, not all of the functionality in a server is helpful. As a result, genetic algorithm choice of features is used to identify the best set of features.

A. Simple Genetic Algorithm Procedure

Initial Population: A population is a comprehensive set of genotypes. Generally defined genetic algorithms with an initial population which is generated randomly.

Fitness-Based Selection: Each chromosome has a selection chance that is directly proportional to its fitness in this type of parent selection.

Reproduction: The steady-state approach selects two chromosomes and crosses them to obtain one or two children, possibly also applies mutation and returns the result to that population.

Crossover Operator: This incorporates data from two genotypes of parents into one or two genotypes of offspring.

Mutation: Mutation has the effect of making it possible to reach all future chromosomes. The mutation operator selects and adjusts every bit position in a string randomly.

Steps in the reduction of dimensions:

1. The preprocessed data includes a number of chromosomes

2. For each individual chromosome, the intensity and informativeness of privacy are determined. Use the mixture of chromosomes to achieve crossover and mutation

3. Evaluate the chromosome newly generated.

4. Repeat the process until a chromosome with maximum privacy and information is obtained.

5. The reduced collection of attributes is generated according to the function

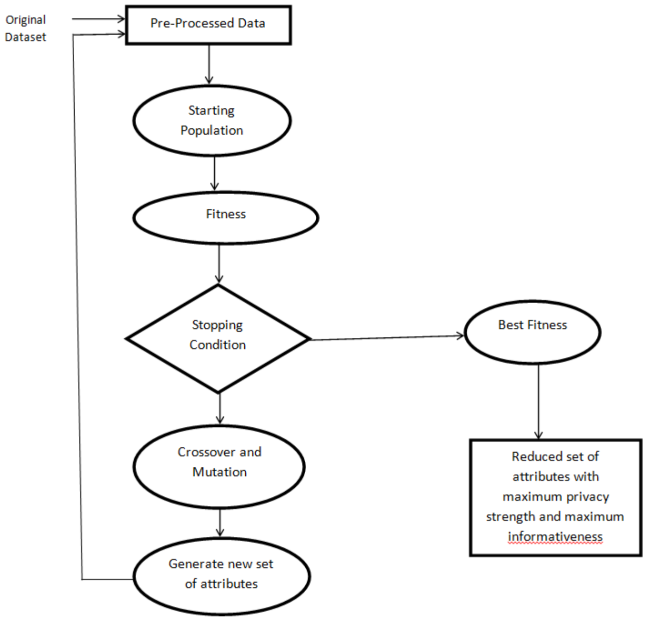

Original Dataset
Pre-Processed Data
Starting Population
Fitness
Stopping Condition
Best Fitness
Crossover and Mutation
Reduced set of attributes with maximum privacy strength and maximum informativeness
Generate new set of attributes

Figure 1: Dimensionality Reduction using Feature Selection

6. Selection process: Sorting based on Gray encoding is achieved by the reduced array of attributes.

7. The sorted transactions form an anonymous group using high-dimensional information correlation-aware anonymization [11].

V. CONCLUSION

The need to protect personal identification details is almost always hindered by the exchange of sensitive data with investigators, but little attention has been given to the study and evaluation of existing data anonymisation systems for data leakage and other performance characteristics. One advantage is that the number of records in the anonymized table is accurate and may be useful in certain applications. It does not help to achieve k-anonymity and eliminating an attribute only eliminates diversity that does not help achieve L-diversity. Eliminating an outlier in t-closeness will smooth a distribution and bring it closer to the overall distribution. To achieve better data quality, these methods should be combined together with generalization and suppression. Generalization cannot handle high-dimensional data, reducing the usefulness of data. Suppression reduces the data quality. Slicing technique involves the horizontal and vertical partitioning of data. Slicing provides efficient data usefulness compared to generalization and is more efficient in comparing bucketization in workload slicing. The benefit of using slicing as a slicing privacy technique is that it is capable of handling high-dimensional data.

VI. REFERENCES

[1].Can Eyupoglu, Muhammed Ali Aydin, Abdul Halim Zaim, and Ahmet Sertbas(2018), “An Efficient Big Data Anonymization AlgorithmBased on Chaos and Perturbation Techniques”.

[2].Dhamodran.P, Priyadharsini.P, Kavitha.M.S(2014), “A Survey On Data Anonymization Techniques For Large Data Sets”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.11.

[3].Johny Antony P, Dr. Antony Selvadoss Thanamani(2017), “Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data”, Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, pp. 247-253.

[4].Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian(2007), “t-Closeness: Privacy Beyond k-Anonymity and L-Diversity”, ICDE.

[5].Mayil Vel Kumar P, Karthikeyan M(2012), “L Diversity on K-Anonymity with External Database for improving Privacy Preserving Data Publishing”, International Journal of Computer Applications (0975 – 8887).

[6].Nithya M and Sheela T(2014), “A Comparative Study on Privacy Preserving DataMining Techniques”, International Journal of Modern Engineering Research (IJMER), Vol 4, Issue 7, ISSN 2249.

[7].Priyank Jain, Manasi Gyanchandani and Nilay Khare(2016), “Big data privacy: a technological perspective and review”, Journal of Big Data, Springer.

[8].Sweeney Latanya Datafly(2016), “A system for providing Anonymity in Medical data”, ACM.

[9].Logeswari A, Thirukumar K (2018), “Privacy preserving data publishing using slicingwith marginal publication”, IJERT.

[10].H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar(2003), “On the privacy preserving properties of random data perturbation techniques,” in ICDM ’03, Melbourne, FL, USA.

[11].Saranya U,Thirukumar K(2018), “Anonymization of high-dimensional data by dimensionality reduction using feature selection”, IJERT.

[12].Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham(2009), “Privacy Preserving Decision Tree Mining from Perturbed Data”, Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1-10.