I.J. Computer Network and Information Security, **2014, 6, 9-17**
Published Online May 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2014.06.02

Improved Anonymization Algorithms for Hiding Sensitive Information in Hybrid Information System

Geetha Mary A
SCSE, VIT University, Vellore, Tamil Nadu, India
Email: geethamary.a@gmail.com

D.P. Acharjya and N. Ch. S. N. Iyengar
SCSE, VIT University, Vellore, Tamil Nadu, India
Email: {dpacharjya, nchsniyengar48}@gmail.com

***Abstract*—**In this modern era of computing, information technology revolution has brought drastic changes in the way data are collected for knowledge mining. The data thus collected are of value when it provides relevant knowledge pertaining to the interest of an organization. Therefore, the real challenge lies in converting high dimensional data into knowledge and to use it for the development of the organization. The data that is collected are generally released on internet for research purposes after hiding sensitive information in it and therefore, privacy preservation becomes the key factor for any organization to safeguard the internal data and also the sensitive information. Much research has been carried out in this regard and one of the earliest is the removal of identifiers. However, the chances of re-identification are very high. Techniques like k-anonymity and l-diversity helps in making the records unidentifiable in their own way, but, these techniques are not fully shielded against attacks. In order to overcome the drawbacks of these techniques, we have proposed improved versions of anonymization algorithms. The result analysis show that the proposed algorithms are better when compared to existing algorithms.

***Index Terms*—**K-anonymity, l-diversity, Data Publication, Anonymization, Privacy preservation, Generalization, Suppression, Datafly algorithm.

I. Introduction

Across the world, quintillions and trillions of data are fabricated and generated from various sources such as blogs, sms, social networks, email etc., but, majority of these are unstructured and it is tricky to pull out knowledge from it. Business intelligence software's focus only on business intension and it's not of much help. Big data analysis is put forward to bail out these problems. It handles multiple sources and reacts fast according to customer requirements. Therefore, organizations keep on collecting data from multiple resources. Some organizations go for publishing these data in World Wide Web for research purposes. On the contrary, Health Insurance Portability and Accountability Act, HIPPA[1], signed by President of USA states exclusively about technical safe guard of patients data. The patient safety and quality improvement act of 2005 ensures patient safety and confidentiality of data but allows limited disclosure. During March and April of 2011, 8.3 and 10 million people were affected by security breaches. Every three months once privacy attacks takes place and personal health information gets stolen. Recently many operations are being outsourced from USA to other countries like India, while doing so confidentiality need to be maintained towards patient details and therefore, this gives a wakeup call for privacy algorithms in order to adhere to the privacy laws enforced in different countries. Since a lot of projects have been outsourced from USA, the data given outside the country should undergo privacy preservation protocols. In India there is no specific act of privacy preservation, however, an individual has right to file a petition in human rights law for privacy disclosure [2]. In addition, privacy preservation is also an important challenge for wireless health care system [3].

Organizations follow some actions to limit disclosure of data while they go for publishing; they try to publish the data in such a manner that the sensitive information such as turnover of an organization, salary of an employee or diseases of a person should not be identified. In general to achieve this, personal identifiers like name, social security number, employee identity, voter's identity and hospital identity are eliminated from the published data. However, there is a possibility to re-identify an individual by using some background information and data linkage techniques [4]. This limitation can be avoided by using privacy preserving data mining techniques (PPDM). Usually data is either distorted or generalized in PPDM, but the main decisive factor is the level to which the data is to be distorted or generalized so that there is no extensive change in exactness of data or the knowledge developed from it.

Privacy preserving data mining (PPDM) is broadly classified into two categories such as input privacy and

output privacy [5]. It is based on the level the analyst wants to hide the data. In input privacy the dataset is initially distorted whereas in output privacy a specific rule is distorted or generalized. One of the output privacy methods is association rule hiding. Some of the input privacy methods are perturbation [6, 7, 8], anonymization [9, 10, 11], multi party computation [12], blocking [13], swapping [14] etc. Perturbation and multi party computation are generally used for quantitative data whereas anonymization and association rule hiding works well over both qualitative and quantitative data. Recent algorithms for anonymization are datafly [10], μ-argus [9], l-diversity [15].

There is a possibility of re-identification of individual objects with the existing algorithms. To overcome these limitations, we have proposed improved versions of the above said algorithms for hiding sensitive information. In addition, we compare the improved versions of the algorithms with the existing algorithms. Rest of the paper is organized as follows: Section 2 provides an overview of the hybrid information system. Foundations of various anonymization algorithms and their drawbacks are discussed section 3. Improved versions of various anonymization algorithms are proposed in section 4. Comparative analysis of the proposed algorithms with the existing algorithms is carried out in section 5. Finally we conclude the paper in section 6.

II. HYBRID INFORMATION SYSTEM

The basic objective of inductive learning and data mining is to learn the knowledge for classification. However, in real world problems, we may not face with simply classification. One such problem is the ordering of objects. On the contrary, we are interested in hiding sensitive information that is present in an information system. Before we discuss, various privacy preservation techniques in order to hide sensitive information, one must know about an information system.

An information system contains a finite set of objects typically represented by their values on a finite set of attributes. Such information system may be conveniently described in a tabular form in which each row represents an object and column represents an attribute. Each cell of the information system contains an attribute value. Now, we define formally an information system as below.

An information system is defined as a quadruple $I = (U, A, V_a, f_a)$ where U is a finite nonempty set of objects called the universe, A is a finite nonempty set of attributes, V_a is a nonempty set of values for $a \in A$, $f_a : U \to V_a$ is an information function. For example, consider a sample information system as presented in Table 1 in which $U = \{o_1, o_2, o_3, o_4, o_5\}$ represents a nonempty finite set of objects; and A = {CGPA, Programme, Year of joining, Year of passing } be a finite set of attributes. If the attribute values are discrete or categorical, we call the information system as qualitative information system. On the contrary, if all the attribute values are non categorical (numerical), we call the information system as quantitative information system. Information system in which the attribute values are either qualitative or quantitative is known as hybrid information system. The information system presented in Table 1 is a good example of hybrid information system, in which the attribute values are either qualitative or quantitative.

Table 1. Hybrid information system

Object	CGPA	Programme	Year of joining	Year of passing	IQ
o_1	8.6	B. Tech.	2004	2008	High
o_2	7.8	M. Tech.	2006	2008	Low
o_3	8.2	M. Tech.	2005	2007	Average
o_4	9.3	B. Tech.	2005	2009	Average
o_5	6.9	B. Tech.	2003	2007	High

III. FOUNDATIONS OF ANONYMIZATION ALGORITHMS

Anonymization, is a term explicated in oxford dictionary as ‘unknown’. Anonymization makes an object indifferent from other objects. In general, distortion and generalization are used to achieve anonymized data set. For sample, consider a hospital information system as presented in Table 2. According to anonymization, identifiers are generally removed from the information system before publishing the information system. One form of anoonymized information system of Table 2 is presented in Table 3.

Table 2. Sample information system

ID	Name	Age	Zip code	Location	Disease
1	Tim	5	632014	Chennai	Osteoporosis
2	John	14	632015	Bangalore	Osteoarthritis
3	Sam	23	632014	Chennai	Osteoporosis
4	Jane	8	632019	Mysore	Osteoporosis
5	Lincoln	45	650302	Cochin	Bronchitis
6	Angelina	42	650312	Trivandrum	Bronchitis
7	Ted	48	650324	Hyderabad	Flu
8	Steve	45	650338	Nellore	Sinusitis

Table 3. Information system after removing identifiers

Age	Zip code	Location	Disease
5	632014	Chennai	Osteoporosis
14	632015	Bangalore	Osteoarthritis
23	632014	Chennai	Osteoporosis
8	632019	Mysore	Osteoporosis
45	650302	Cochin	Bronchitis
42	650312	Trivandrum	Bronchitis
48	650324	Hyderabad	Flu
45	650338	Nellore	Sinusitis

The data presented in Table 3 is not fully anonymized since individual person could be re-identified from the

table. If a person knows the ZIP from which the patient belongs, the province of the person and age of the person, then the individual patient disease could be identified. Using background knowledge of a patient a user can extract knowledge from the information system by linking the attribute values. For example, from Table 3 it is clear that the patient residing in zip 632015 is suffering from osteoarthritis. Likewise if the patient age is around 5 and from Chennai then the patient has visited hospital for having Osteoporosis. It indicates that, the information system that has been published without applying efficient algorithms is prone to identification. To avoid such limitation, generalization of attribute values in a distorted way is introduced and we call it as suppression, which is discussed in next section. There are three major methods generally employed to achieve anonymization are k-anonymization, l-diversity and t-closeness.

A. Identifiers, Generalization and Suppression

This section unfolds the notions and concepts associated in anonymization algorithm. Identifier attribute is an attribute using which an individual's record is identified. For example, employee identity, patient identity, name, phone number etc are considered as identifiers. Another important attribute in anonymization algorithm is sensitive attribute. The data which a user wants to hide from other individuals is known as sensitive attribute. For example, salary of an employee, patients affected by a disease etc are sensitive attributes.

Distinctive attributes other than identifier attribute and sensitive attribute and the combination of the attributes using which a record could be uniquely re-identified are called as quasi identifiers. Let A be the set of attributes $(a_1, a_2, ..., a_k)$ with sensitive attributes *(SI)*, $(a_i, a_{i+1}, ..., a_j)$; $1 \le i, j \le k$ and identifier attributes *(IA)*, $(a_p, a_{p+1}, ..., a_q)$; $1 \le p, q \le k$. The quasi identifier *(QI)*, are subset of attributes *A*, such that *QI* is neither a subset of SI nor a subset of *IA*. For example in Table 2, using the combination of age, zip and location, a patient could be identified.

Usually the data is published by removing the identifiers and generalizing the values. After removing the identifiers, i.e., identity (ID) and name from Table 2, information systems are generally published like Table 4, after generalization of the attribute values. Suppression is a generalization of values in a distorted way. The operation used in Table 4 on ZIP is suppression. The other operation done on location date is called as generalization. In a hierarchy, moving up a hierarchy is generalization. Suppression is a form of generalization and is explained in Fig. 1, 2, 3, 4 and 5. A representation of generalization and suppression of attribute values of zip is presented in Fig. 1 and 2 respectively, which doesn't have much difference since applied over numerical values. But while handling discrete values, lot of variation is found among suppressing and distorting the values which can be inferred from Fig. 3 and 4. The impact of applying both generalization and suppression is presented in Fig. 5. Attribute values of age is generalized, zip is suppressed and location is generalized in Table 4.

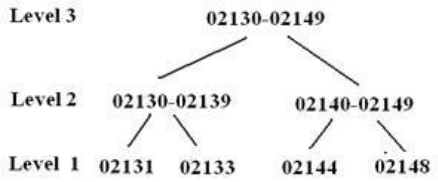

Fig 1. Generalization of the attribute ZIP

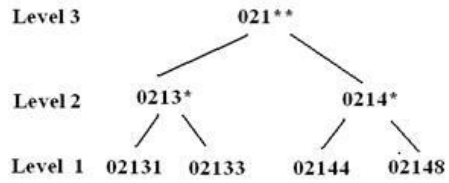

Fig 2. Suppression of the attribute ZIP

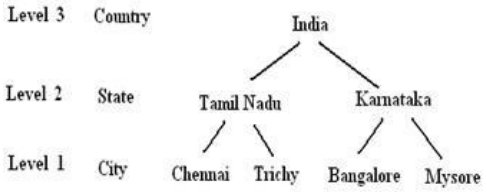

Fig 3. Generalization of the attribute location

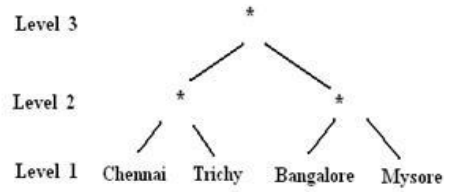

Fig 4. Suppression of the attribute location

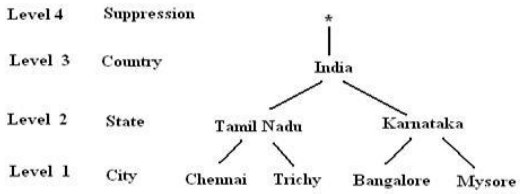

Fig 5. Generalization with Suppression of the attribute location

Table 4. Anonymized information system

Age	Zip	Location	Disease
5-23	63201*	Tamil Nadu	Osteoporosis
5-23	63201*	Karnataka	Osteoarthritis
5-23	63201*	Tamil Nadu	Osteoporosis
5-23	63201*	Karnataka	Osteoporosis
42-48	6503**	Kerala	Bronchitis
42-48	6503**	Kerala	Bronchitis
42-48	6503**	Andhra Pradesh	Flu
42-48	6503**	Andhra Pradesh	Sinusitis

B. K-Anonymization

The level of generalization is set based upon the user requirement of threshold of anonymity. The threshold is set up by a *k*-constant where the value of '*k*' depends upon the minimum level of anonymization required by the user. Let *k* be an integer. After considering the value *k*, the attribute values are generalized in such a way that at least *k* number of attribute values will remain identical. As a result, the probability of identifying an object will be less as compared to simple anonymization. Table 4 presented earlier is a 2-anonymization of the hospital information system. *K*-Anonymization [9, 10, 11] is done using different algorithms like datafly, μ-argus, MinGen. However, it has some limitations as it is vulnerable to some of the attacks like homogeneity attack and background knowledge attack.

Table 5. 2-Anonymous information system

Age	Zip	Location	Disease
<30	63201*	Tamil Nadu	Osteoporosis
<30	63201*	Karnataka	Osteoarthritis
<30	63201*	Tamil Nadu	Osteoporosis
<30	63201*	Karnataka	Osteoporosis
4*	6503**	Kerala	Bronchitis
4*	6503**	Kerala	Bronchitis
4*	6503**	Andhra Pradesh	Flu
4*	6503**	Andhra Pradesh	Sinusitis

Consider a k-anonymization information system presented in Table 5. It is clear that, 1st and 3rd objects are anonymous, since the attribute values in the those objects are same. If the user knows that the object (patient) is less than 30 years of age and is from Tamil Nadu, then the user is 100% sure that the patient has Osteoporosis. Such type of attacks is known as homogeneity attack. From Table 5, it is clear that 7th object is indiscernible with the 8th object though the information system is two anonymous. Consider the user knows that the patient is around 40 years of age and from Andhra Pradesh. Since, the user knows that the patient is suffering from the disease for a long duration, and then the user can conclude that the object (patient) is affected by sinusitis. In such cases, attacker uses his back ground knowledge to get the information. We call such type of attack as background knowledge attack.

C. l-diversity

Since k-anonymization is susceptible to homogeneity and background knowledge attacks, the next level of anonymization is l-diversity proposed by Ashwin et al. [15]. An information system is l-diverse, if no 'l' objects have same attribute values. However, it is susceptible to similarity and skew-ness attack. Consider the information system as shown in Table 6. It is 2-diverse and does not suffer from homogeneity attack. However, if the attacker knows that the object (patient) is from zip code 63201*, then the attacker knows that the object (patient) suffers from some bone related disease. Since Osteoporosis and Osteoarthritis are some form of bone related diseases which are inferred using Fig. 6. Consider another instance of an information system as shown in Table 6. It is clear that, the Table 6 is 2-diverse, but the possibility of finding out an object (person) suffering from Osteoporosis is 75%. So, even l-diversity has some privacy issues.

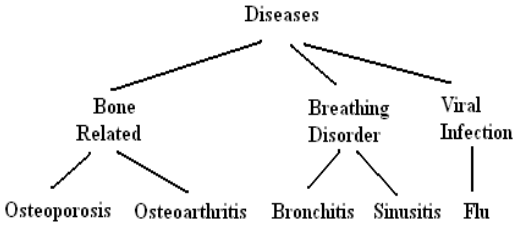

Fig 6. Hierarchy of the diseases

Similarity attack is taken into consideration and proposal of new algorithm is done which is free from the similarity attack. The new algorithm for l-diversity is discussed in next section.

Table 6. 2-diverse information system

Age	Zip	Location	Disease
<30	63201*	*	Osteoporosis
<30	63201*	*	Osteoarthritis
<30	63201*	*	Osteoporosis
<30	63201*	*	Osteoporosis
4*	6503**	*	Bronchitis
4*	6503**	*	Bronchitis
4*	6503**	*	Flu
4*	6503**	*	Sinusitis

D. t-closeness

Since l-diversity suffers from similarity attack, Ninghui [16] proposed t-closeness where the distribution of sensitive attribute in this class and the distribution of the attribute in the information system is not more than the threshold value t. So, the sensitive attribute value is distributed perfectly among the objects and protects from similarity attack. It uses Earth Mover's distance to uphold the threshold value t.

IV. Proposed Algorithms for Hiding Sensitive Information

In this section, we propose improved versions of datafly, µ-argus and l-diversity algorithms proposed by different researchers. Now we formally discuss the improved version of the mentioned algorithms in the following section.

A. Improved Datafly Algorithm

Datafly algorithm discussed by Sweeney [10] uses generalization. However by using both generalization and suppression, the improved version provides better results. Now we formally discuss the improved version of the datafly algorithm. Let I be the initial information system and DMG is the domain generalization. We select the quasi identifiers (QI) as for user's requirement. The data values in the information system are generalized and the distinct data values that are present in the information system are removed to get private information system. The main disadvantage of the existing datafly algorithm is that it reduces the record size in the private information system. It indicates that the records that are present in the information system may not be present in the private information system. In order to overcome this limitation, we perform local suppression on the identifiable distinct values. As a result the records in the information system are retained by generalizing and suppressing the records. The improved version of the algorithm is presented below.

Algorithm1. Improved datafly algorithm

Input: Information system (*I*), Quasi identifier (*QI*) and domain generalization (*DMG*)
Output: Private information system that retains all the objects.

1. Construct frequency list of each quasi identifier
2. Generalize the quasi identifier attribute having highest frequency count
3. Identify the equivalence classes based on the quasi identifiers
4. Compute the cardinality of each equivalence class. We call it as frequency of that equivalence class.
5. If the frequency of the object is less than k (integer), then do suppress the quasi identifiers one by one until we get the frequency of the object greater than or equal to k
6. Else go to step 2
7. Publish the private information system
8. End

B. Improved µ-argus algorithm

Netherland national statistical agency generally use µ-argus algorithm for hiding sensitive information. It starts working by assigning each attribute to any of the categories such as most identifying, not identifying, more identifying and less identifying. It then identifies the outliers by combinations of quasi identifiers. Finally, cell level suppression is carried out. The algorithm has been elaborately discussed by Sweeney [11]. The main limitation of µ-argus algorithm as specified by Sweeney [9] is that it is not completely k-anonymous, therefore µ-argus algorithm need to be further processed in order to achieve k-anonymization for a given information system. The improved µ-argus algorithm overcomes the limitation of existing algorithm.

Algorithm 2. Improved µ-argus algorithm

Input: Information system (I), Quasi identifier (QI) and domain generalization (DMG)
Output: Private information system that retains all the objects.

1. Construct frequency list of each quasi identifier
2. Generalize the quasi identifier attribute having highest frequency count
3. Identify the equivalence classes based on the quasi identifiers
4. Compute the cardinality of each equivalence class. We call it as frequency of that equivalence class.
5. If the frequency of the object is less than k (integer), then do the following
   (a) compute all possible combinations of quasi identifiers
   (b) Identify the possible combinations that uniquely identify an object of an information system. We call it as outliers.
   (c) Identify the quasi identifiers that is repeated maximum number of times in the outliers
   (d) Suppress the quasi identifier identified in the previous step
6. Steps 2, 3, 4 are repeated until we achieve k-anonymization
7. Publish the private information system
8. End

C. Improved l-diversity algorithm

Ashwin et al [15] discussed l-diversity algorithm to hide sensitive information in an information system. It is observed that l-diversity suffers from similarity attack. It means that, the attacker would not be able to identify a particular decision but he could identify the type of decisions that a particular object may have. In order to overcome this limitation we propose improved l-diversity algorithm.

Algorithm 3. Improved l-diversity algorithm

Input: Information system (I), Quasi identifier (QI) and domain generalization (DMG) and Diseases Hierarchy
Output: Private information system that retains all the objects.

1. Construct a sub-tree based on sensitive information. Let each node of the tree belong to a particular category.
2. Generalize the quasi identifier.
3. Identify the equivalence classes based on the quasi identifiers
4. Compute the cardinality of each equivalence class. We call it as frequency of that equivalence class.
5. If the sensitive information of an equivalence class does not belong to l-(integer) different category, then do the following
   (a) Swap the objects of the equivalence class with the objects of different equivalence class.
   (b) Suppress the attribute values of the quasi identifiers in such a manner that their attribute values are similar.
   (c) Do repeat step a to c until the sensitive information of an equivalence class belongs to l-different categories.
6. Publish the private information system
7. End

V. Numerical Illustration of Improved Algorithms

In this section we explain our improved version of algorithms with a numerical illustration. Let us consider an information system with 12 objects as shown in Table 7, where the attributes are defined based on the problem objective. Before we impose improved version of algorithms, we compute the reduct, data cleaning and preprocessing, in order to minimize the computational complexity. In particular we remove the objects that contain missing attribute values. In the following section we analyze the proposed algorithms based on the information system presented in Table 7.

Table 7. Information system for numerical illustration

ID	Name	Race	Date of birth	Gender	ZIP	Problem
1	Tim	black	9/20/1965	male	2141	Asthma
2	Sam	black	2/14/1965	male	2141	Pulmonary vascular
3	Ana	black	10/3/1965	female	2138	Lung cancer
4	Andrea	black	8/24/1965	female	2138	Mouth cancer
5	Louis	black	11/7/1964	female	2138	Coronary artery
6	Rose	black	12/1/1964	female	2138	Lung cancer
7	Steve	white	9/23/1964	male	2138	Pulmonary vascular
8	Angel	white	3/15/1965	female	2139	Cardio myopathy
9	Tom	white	8/13/1964	male	2139	Coronary artery
10	Ben	white	5/5/1964	male	2139	Lung cancer
11	Joe	white	2/13/1967	male	2138	Pulmonary vascular
12	Kim	white	3/21/1967	male	2138	Asthma

A. Analysis of Improved Datafly Algorithm

In this section we analyze the improved version of datafly algorithm presented in section 4 with the help of information system stated in Table 7. Initially from the attributes quasi identifiers are identified and generalized. Further the equivalence classes based on the quasi identifiers are computed. Cardinality of each equivalence class is computed and is presented as frequency of that equivalence class as shown in Table 8. From Table 8 it is clear that the frequency of the equivalence class with combination race - white, date of birth - 1964, gender - male, zip - 2138 is 1. On considering 2-anonymization, the equivalence class is removed from the information system as mentioned in the existing algorithms [10]. Similarly the other equivalence classes having frequency less than 2 are removed from the information system before publishing the information system.

Table 8. Frequency computation of information system

Race	Date of birth	Gender	Zip	Frequency
black	1965	male	2141	2
black	1965	female	2138	2
black	1964	female	2138	2
white	1964	male	2138	1
white	1965	female	2139	1
white	1964	male	2139	2

The output of the existing datafly algorithm is presented in Table 9. From Table 9 it is clear that the number of objects in an information system is reduced after imposing the existing datafly algorithm. In the proposed algorithm, instead of removing the object from the information system we suppress the quasi identifiers of that concerned objects in such a manner that we achieve k-anonymization. The result on imposing the improved version of the datafly algorithm, we publish the private information system as shown in Table 10. The main advantage of the improved version of the algorithm is that it never reduces the number of objects in an information system.

Table 9. 2-anonymous information system generated by datafly algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	male	2141	Asthma
black	1965	male	2141	Pulmonary vascular
black	1965	female	2138	Lung cancer
black	1965	female	2138	Mouth cancer
black	1964	female	2138	Coronary artery
black	1964	female	2138	Lung cancer
white	1964	male	2139	Coronary artery
white	1964	male	2139	Lung cancer
white	1967	male	2138	Pulmonary vascular
white	1967	male	2138	Asthma

Table 10. 2-anonymous information system generated by improved datafly algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	male	2141	Asthma
black	1965	male	2141	Pulmonary vascular
black	1965	female	2138	Lung cancer
black	1965	female	2138	Mouth cancer
black	1964	female	2138	Coronary artery
black	1964	female	2138	Lung cancer
white	196*	male	2138	Pulmonary vascular
white	196*	female	2139	Cardiomyopathy
white	1964	male	2139	Coronary artery
white	1964	male	2139	Lung cancer
white	1967	male	2138	Pulmonary vascular
white	1967	male	2138	Asthma

B. Analysis of Improved μ-argus Algorithm

This section analyses the improved version of μ-argus algorithm presented in section 4 with the help of information system stated in Table 7. Initially from the attributes quasi identifiers are identified and generalized. Further the equivalence classes based on the quasi identifiers are computed. Cardinality of each equivalence class is computed and is presented as frequency of that equivalence class as shown in Table 8. From Table 8 the outliers are identified for each object in the equivalence class whose frequency is less than k. In the present instance we have considered k=2 and the outliers produced by existing μ-argus algorithm is presented in Table 11. From Table 11 it is clear that the occurrence of the quasi identifier, date of birth is maximum and so it is considered as the most identifying quasi identifier. So the combinations are generated based on the existing algorithm [11] and the values are suppressed accordingly. The analysis of the existing algorithm is presented in Table 12. However it is observed that the private information system presented in Table 12 is not completely 2-anonymous. In order to overcome this limitation we identified that the next frequent quasi identifier and suppress them based on their values. The process is repeated until we achieve k-anonymization. The output of the improved version of the algorithm is presented in Table 13. From Table 13 it is clear that the improved version is completely 2-anonymous.

Table 11. Outliers generated for μ-argus algorithm

S.No	Outliers		
1	Race	Date of birth	Zip
2	Gender	Date of birth	Zip
3	Race	BirthDate	Gender
4	Race	BirthDate	
5	Race	Gender	
6	Zip	BirthDate	
7	Gender	Zip	

Table 12. 2-anonymous information system generated by μ-argus algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	male	2141	Asthma
black	1965	male	2141	Pulmonary vascular
black	1965	female	2138	Lung cancer
black	1965	female	2138	Mouth cancer
black	1964	female	2138	Coronary artery
black	1964	female	2138	Lung cancer
white	1964	male	2138	Pulmonary vascular
white		female	2139	Cardiomyopathy
white	1964	male	2139	Coronary artery
white	1964	male	2139	Lung cancer
white	1967	male	2138	Pulmonary vascular
white	1967	male	2138	Asthma

Table 13. 2-anonymous information system generated by improved μ-argus algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	male	2141	Asthma
black	1965	male	2141	Pulmonary vascular
black	1965	female	2138	Lung cancer
black	1965	female	2138	Mouth cancer
black	1964	female	2138	Coronary artery
black	1964	female	2138	Lung cancer
white		male		Pulmonary vascular
				Cardiomyopathy
white	1964	male	2139	Coronary artery
white	1964	male	2139	Lung cancer
white	1967	male	2138	Pulmonary vascular
white	1967	male	2138	Asthma

C. Analysis of Improved l-diversity algorithm

The improved version of l-diversity algorithm as discussed in section 4 is analysed in this section. Initially from the attributes, quasi identifiers are identified and generalized. Further the equivalence classes based on the quasi identifiers are computed. In l-diversity algorithm the sensitive information, i.e., disease is categorized into to various categories as shown in Fig. 7. It is observed that, the equivalance classes obtained based on quasi identifiers belong to a particular category and face similarity attack. This is clearly identified from Table 14. In addition, from Table 14 it is identified that the objects of the first equivalence class satisfying k-anonymization is having some respiratory diseases whereas the objects of second equivalence class satisfying k-anonymization suffers from cancer related diseases. In order to overcome this limitation we swapped the objects of one equivalence class to another based on the disease hierarchy. Further the quasi identifier values are suppressed. Table 15 presents the output of the improved version of the algorithm. From Table 15 it is clearly identified that the private information system is free from similarity attack.

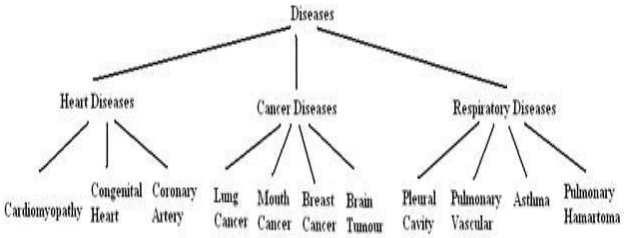

Fig 7. Hierarchy of the diseases

Table 14. 2-diverse information system generated by l-diversity algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	*	21*	Asthma
black	1965	*	21*	Pulmonary vascular
black	1965	*	21*	Lung cancer
black	*	female	2138	Mouth cancer
black	*	female	2138	Coronary artery
black	*	female	2138	Lung cancer
white	196*	*	213*	Pulmonary vascular
white	196*	*	213*	Cardiomyopathy
white	196*	*	213*	Coronary artery
white	196*	male	213*	Lung cancer
white	196*	male	213*	Pulmonary vascular
white	196*	male	213*	Asthma

Table 15. 2-diverse information system generated by improved l-diversity algorithm

Race	Date of birth	Gender	Zip	Problem
black	1965	*	021*	Asthma
black	1965	*	021*	Pulmonary vascular
black	1965	*	021*	Lung cancer
black	196*	female	02138	Mouth cancer
black	196*	female	02138	Coronary artery
black	196*	female	02138	Lung cancer
white	196*	*	02138	Pulmonary vascular
white	196*	*	02139	Cardiomyopathy
white	196*	*	02139	Coronary artery
white	196*	male	0213*	Lung cancer
white	196*	male	0213*	Pulmonary vascular
white	196*	male	0213*	Asthma

VI. Conclusion

Hiding sensitive information in hybrid information system is very important nowadays. In the recent algorithms developed for anonymization by various researchers such as datafly, µ-argus and l-diversity, there is a possibility of re-identification of individual objects. In this paper we have proposed improved version of the existing algorithms in order to overcome its limitations. We have taken a numerical example in order to analyze the improved versions of the algorithms with the existing versions. The results obtained for the improved versions are identified as better compared with existing algorithms. The results obtained also show the practical viability of the proposed research.

References

[1] U.S. Department of Health and Human Services, Health Insurance Portability and Accountability Act (2006). Summary of HIPPA privacy rule [online]. U.S. Department of Health & Human Services, Washington, D.C. (Accessed 23 October 2013).

[2] Shrikant, A., Tanu, S., Swati, S., Vijay, C. and Abhishek, V., Privacy and Data Protection in Cyberspace in Indian Environment, International Journal of Engineering Science and Technology, 2010, 2(5), p. 942-951.

[3] Qiming H.,Xing Y.,Shuang Li, Identity Authentication and Context Privacy Preservation in Wireless Health Monitoring System, International Journal of Computer Network and Information Security, 2011, Vol. 3, No. 4, p. 53-60.

[4] Peter, C., A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication, IEEE transactions on knowledge and data engineering, 2012, 2(9), p.1537-1555.

[5] Wang, S. L. and Jafari, A., Hiding sensitive predictive association rules, In proceedings of IEEE International Conference on Systems, Man and Cybernetics, Proceedings, 2005, 1, p. 164-169.

[6] Rakesh, A. and Ramakrishnan, S., Privacy-Preserving Data Mining, In 2000 ACM SIGMOD International Conference on Management of Data, New York, USA, Proceedings, 2000, p. 439-450.

[7] Agrawal, D. and Aggarwal, C., On the design and quantification of privacy preserving data mining algorithms, In the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, New York, USA, Proceedings, 2001, p.247-255.

[8] Kargupta, H., Datta, S. ,Wang, Q. and Krishnamoorthy, S., On the privacy preserving properties of random data perturbation techniques, In Proceedings of the ICDM 2003- 3rd IEEE International Conference on Data Mining, Los Alamitos, California, 2003, p. 99-106.

[9] Sweeney, L., Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5), p.571-588.

[10] Sweeney, L., Guaranteeing anonymity when sharing medical data the Datafly system, In Journal of the American Medical Informatics Association, Washington, DC: AMIA, Proceedings, 1997, pp. 51-55.

[11] Sweeney, L., Computational disclosure control: A primer on data privacy protection, Ph.D. Thesis, Massachusetts Institute of Technology, 2001.

[12] Lindell, Y. and Benny, P., Secure Multiparty Computation for Privacy Preserving Data Mining, The Journal of Privacy and Confidentiality, 2009, 1(1), p. 59-98.

[13] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y. and Theodoridis, Y., State-of-the-art in privacy preserving data mining, SIGMOD Rec, 2004, 33(1), p. 50-57.

[14] Moore, R. A. Jr., Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets, Statistical Research Division Report Series RR 96-04, US Bureau of the Census, 1996.

[15] Ashwin, M., Johannes, G. and Danieal, K., l-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1), p. 1-52.

[16] Ninghui, L., Tiancheng, L. and Venkatasubramanian, S., t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, In ICDE 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, Proceedings, 2007, p.106-115.

Geetha Mary A, received her M. Tech. in computer science and engineering from VIT university, Vellore, India in the year 2008 and B.E. from University of Madras, Tamil Nadu,India in the year 2004. She is currently working for VIT University as Assistant Professor-Senior and is pursuing her Ph. D. at VIT University. Her research interests include Security for Data Mining, Databases and Intelligent Systems.

D. P. Acharjya, received M. Sc. from NIT, Rourkela, India; M. Tech. in computer science from Utkal University, India and obtained his Ph. D. from Berhampur University, India. He has been awarded with Gold Medal in M. Sc. He is presently working as a Professor in the School of Computing Sciences and Engineering at VIT University, India. He has authored many international and national papers and four books to his credit. In addition, he has edited two books with IGI Global, USA. Dr. Acharjya is associated with many professional bodies CSI, ISTE, IMS, AMTI, ISIAM, OITS, IACSIT, CSTA and IAENG.

N. Ch. Sriman Narayana Iyengar, received Ph. D from Regional Engineering College Warangal, Kakatiya University, Andhra Pradesh, India. He is working as a Senior Professor at the School Of Computing Science and Engineering at VIT University, Vellore, India. His research interests include fluid dynamics (Porus Media), agent based e-business applications, data privacy, information security, mobile commerce and cryptography. He has authored several textbooks and had research publications in national, international journals and conferences. He is editorial board member for many national and international Journals.

How to cite this paper: Geetha Mary A, D.P. Acharjya and N. Ch. S. N. Iyengar,"Improved Anonymization Algorithms for Hiding Sensitive Information in Hybrid Information System", IJCNIS, vol.6, no.6, pp.9-17, 2014. DOI: 10.5815/ijcnis.2014.06.02

