Privacy and Utility Preserving Task Independent Data Mining

E. Poovammal
Department of CSE
SRM University - 603203
Chennai, India

M. Ponnavaikko
Bharathidasan University
Trichi
Tamilnadu, India

ABSTRACT
Today's world of universal data exchange, there is a need to manage the risk of unintended information disclosure. Publishing the data about the individuals, without revealing sensitive information about them is an important problem. K-anonymization is the popular approach used for data publishing. The limitations of K- anonymity were overcome by methods like L-diversity, T-closeness, (alpha, K) anonymity; but all of these methods focus on universal approach that exerts the same amount of privacy preservation for all persons against linking attack, which result in high loss of information. Privacy was also not guaranteed 100% because of proximity and divergence attack. Our approach is to design micro data sanitization technique to preserve privacy against proximity and divergence attack and also to preserve the utility of the data for any type of mining task. The proposed approach, apply a graded grouping transformation on numerical sensitive attribute and a mapping table based transformation on categorical sensitive attribute. We conduct experiments on adult data set and compare the results of original and transformed table to show that the proposed task independent technique preserves privacy, information and utility.

Categories and Subject Descriptors
H.2.7 [Database Management]: Database Administration| Security, integrity, and protection; H.2.8 [Database Management]: Database Applications|, Data mining

General Terms
Algorithms, Performance, Security

Keywords
Anonymization, Data Publishing, Data utility, Privacy management.

1. INTRODUCTION
Micro data (table with un-aggregated data) is a valuable source of information for research. However, disclosure of individuals' private sensitive information is unacceptable. Micro data is stored and released by a trusted centralized server. To avoid identification of individuals' information, attributes like Name, SSN (Social Security Number) are removed from the table before release. But, some set of attributes called Quasi Identifiers (QI), for example, Age, Gender and Zip can be linked with external data to uniquely identify the individual. The association of quasi-identifiers with sensitive attributes in public records is known as sensitive attribute disclosure. This problem is named as linking attack. It is very easy to prevent sensitive attribute disclosure by simply not publishing Quasi Identifiers and sensitive attributes together. But, the only reason to publish generalized Quasi Identifiers and sensitive attributes together is to support data-mining tasks that consider both types of attributes in the sanitized database.

To counter the linking attack problem, K- anonymity was the first and effective approach applied [22][23]. K – anonymity can not be applied to high dimensional data without complete loss of utility [3']. Since K-anonymity does not put any restriction on sensitive attributes, it faces homogeneity attack and background knowledge attack [16]. In other words, K-anonymity creates groups which leak information due to lack of diversity within the group. This limitation is overcome by L-diversity principle [16]. But, L-diversity principle deals with diversity only in categorical sensitive attributes (e.g. Disease) of a QI group. It does not check for range of values in numerical sensitive attribute (e.g. Income) and hence leads to the problem of proximity breach.

Proximity breach occurs in numerical sensitive attribute, when an adversary concludes with high confidence that the sensitive value of a victim must fall in a short interval even though the adversary may have low confidence about the victim's actual value [12].

Utility of any data set can be measured only if the computation (task) to be performed on that data set is known. For example, a classification task performed on a real dataset may yield better results than that of the clustering task performed on the same data set. So, the sanitization should be in such a way that any task performed on the original and the sanitized data should yield almost similar results. Generalization of Quasi Identifiers by suppression [23] is also a sanitization method which leads to information loss. Since, many data mining tools can not handle missing attributes; the utility of the data gets reduced.

Without knowing the task to be applied to the data set, it is meaningless to judge about the utility of the data. But, the main objective of privacy preserving techniques is to preserve good utility of data, independent of the task applied. The proposed method maintains the structure and data type of the sensitive attribute while hiding the actual values. So, the utility of the data is as good as the original data.

1.1 Motivation

We identify *Divergence breach* as a privacy threat specific to categorical sensitive attributes in anonymized data publication. Divergence breach occurs in categorical sensitive attribute, when an adversary concludes with high confidence that the sensitive value of the victim individual is completely irrelevant to the actual value, even though the adversary may have low confidence about the victim's actual value.

Both the proximity and divergence attack exist in a published table, as long as there exist a QI group. The main issue in releasing the micro data is the disclosure of sensitive information. Surprisingly, in the literature almost all techniques deal with quasi identifier attributes or with sensitive attributes within a QI- group. The proposed technique deals with sensitive information directly, without disturbing QI attribute values there by preserving the information. We replace the original sensitive data with new data, which exhibit same general pattern but conceal the sensitive information even if traced by linking attack.

1.2 Our Contributions

First, we explain the problem of divergence breach in a K-anonymized, L-diverse table.

Second, the methodology of transformation of sensitive attributes is elaborated. A graded grouping approach is chosen for numerical sensitive attribute and a mapping approach is chosen for categorical sensitive attribute. All the sensitive attributes are to be transformed accordingly to get the transformed table. No other operation needs to be done on the quasi identifier or neutral attributes.

Third, it is explained that the above transformation procedures preserve the properties of the original data thereby information is maintained without any privacy leakage.

Fourth, to evaluate our methodology different mining tasks are performed on the original Adult dataset from the UCI machine learning repository. Then, the same types of tasks are again performed on the transformed table. The results are compared depending on the task.

2. RELATED WORK

The issue of privacy preservation was handled by many ways like cryptographic methods in distributed environment [24] [21], changing the results of application and algorithms [6] [20], query answering and data publishing. Query answering techniques are much related to data publishing approach, where instead of publishing the data, the data base answers queries as long as the answers do not breach privacy. Query restriction, auditing and output perturbation techniques and other statistical approaches are discussed in [1]. Auditing [13] and output perturbation [10] require maintaining state about the previous queries while data publishing does not need to maintain any state of queries asked.

The data publishing techniques tend to perform data transformation to maintain privacy. These techniques include methods such as data perturbation [5] [4] [25], K-anonymity [22] [23] [7] [14], L-diversity [16] and T-closeness [19]. The data perturbation is a technique of adding/multiplying noise to the original data. Instead of original table, only the perturbed table and the distribution function used for perturbation are released for analysis. It is impossible to reconstruct the original data set and also the accuracy level is sensitive to the reconstruction algorithm [4]. If anonymization is done by suppression method, in addition to information loss, the utility of the data gets reduced. Because many mining tools and algorithms do not work on attributes with missing values.

Example: Consider the Patients List shown in Table 1, released by a trusted server of a hospital and the Voters List of that area released shown in Table 2. Table 1 does not contain any uniquely identifying attributes like Name, SSN. The attributes Disease, Income of table 1are sensitive attributes. An attribute is called *Sensitive*, if the individual is not willing to disclose or an adversary must not be allowed to discover the value of that attribute. The attribute Income of Patient table is considered for analyzing his/her spending capacity. The collection of attributes {Age, Gender, and Zip} is called the Quasi Identifier (QI) attributes; Because, by linking the QI attributes of these two tables an adversary gets the value of identifying attribute Name, from the Voters List. So, the sensitive information for example, disease of Barbie (Stomach cancer) and her income (15000) is disclosed.

Table 1. Patients List

Age	Sex	Zip Code	Income	Disease
33	M	600018	22000	Flu
29	F	600008	15000	Stomach Cancer
21	M	600006	10000	Bronchitis
31	M	600009	20000	Gastritis
22	M	600006	10020	Bronchitis
60	M	600019	23000	Flu
25	F	600006	10030	Bronchitis

Table 2. Voters List

Disease	Age	Sex	Zip Code
Anand	33	M	600018
Barbie	29	F	600008
Charles	21	M	600006
Dinesh	31	M	600009
Esra	22	M	600006
Febi	60	M	600019

Girija	25	F	600006

2.1 K-Anonymity

A table satisfies K-anonymity if every record in the table is indistinguishable from at least K-1 other records in a QI- group. In a QI-group all the values of the quasi identifier attributes of K-records are identical. Such a table is called K-anonymous table. The K-anonymity requirement is typically enforced through generalization, where real values of QI attributes are replaced with *less specific but semantically consistent values* [23]. K-anonymity guarantees that an individual can be associated with his/her real tuple with a probability of at most 1/K, through linking attack.

But K-anonymity only prevents association between individuals and tuples instead of association between individuals and their sensitive values. Since, this method places no constraint on the sensitive values in each QI-group; it may result in homogeneity attack. Homogeneity attack allows an adversary to derive the sensitive information of an individual with 100% confidence. Assume, an adversary attempts to infer the disease of Girija knowing her age and Zip. From the published 3- anonymous patient table 3, the adversary knows that she belongs to the QI-group 1, and declares with 100% confidence that Girija is affected with Bronchitis. This problem is the motivation for L-diversity principle [16].

2.2 L-Diversity

A table is L-diverse if every QI-group is L-diverse. A QI group is L-diverse, if contains at least *L-well represented* values for sensitive attribute. Homogeneity attack in the 3-anonymous patient table 3 is prevented by releasing 3-anonymous, 2-diverse patient table 4. The L-diversity principle is improvised in [26][28] and some other different methods like T-closeness [19], (c,k) safety [17].

The choice of the principles depends on the needs of underlying application. But all these methods focus on universal approach that exerts the same amount of preservation for all persons without catering for their concrete needs. But, Justin Brickell [8] prefers publication of QI and Sensitive attribute values directly in two different tables. The only reason to publish quasi identifier attributes and sensitive attributes together is to support data mining tasks that consider both types of attributes in a database. An effective condensation method developed in [2] but releases only selected statistics about each QI group.

A new generalization framework based on concept of personalized privacy which can maintain large amount of information from micro data was presented in [27].

2.3 Personalized Privacy Preservation

Since, K-anonymity and its improved version such as L-diversity can not guarantee privacy protection if an individual corresponds to multiple tuples in the micro data; sensitive attribute (SA) generalization is introduced in [27]. After QI generalization, SA generalization (categorical data) is performed for all the QI groups, based on guarding node set by the individuals. So, sensitive values of each QI group is offered with required amount of privacy protection (varies with respect to each group), there by increasing the information by reducing excessive generalization.

The numerical sensitive attributes are categorized and the taxonomy tree built on this categorization is used for selecting the guard node. But, similar to optimal QI generalization [11], optimal SA generalization is NP hard. Also, SA generalization does not take care of diversified values in a QI- group.

2.4 (e – m) Anonymity

Even if there exists, L-well represented values in sensitive numerical attribute of a QI group in a L-Diverse table; there is a possibility that an adversary, knowing the QI-group, concludes a short range sensitive value of the victim, with high confidence. For example, from the 3-Anonymous 2-Diverse Patient Table 4, an adversary knowing that the victim belongs to first QI group concludes victim's salary is in the interval (10000-10030) with 75% probability, although he has 25% chance to discover the actual salary of the victim. This problem is named as proximity breach and handled by (e-m) anonymity method [12] which demands that, given a QI group G, for every sensitive value x in G, at most 1/m of the tuples in G can have sensitive values similar to x where the similarity is controlled by e.

Table 3. Three-Anonymous Patient Table

Age	Sex	Zip Code	Income	Disease
21-25	Person	600006	10000	Bronchitis
21-25	Person	600006	10020	Bronchitis
21-25	Person	600006	10030	Bronchitis
29-60	Person	600008 – 600019	22000	Flu
29-60	Person	600008 – 600019	15000	Stomach Cancer
29-60	Person	600008 – 600019	20000	Gastritis
29-60	Person	600008 – 600019	23000	Flu

Table 4. Three-Anonymous, Two-Diverse Patient Table

Age	Sex	Zip Code	Income	Disease
21-29	Person	600006 – 600008	10000	Bronchitis
21-29	Person	600006 – 600008	10020	Bronchitis
21-29	Person	600006 – 600008	10030	Bronchitis
21-29	Person	600006 – 600008	15000	Stomach Cancer
31-60	Person	600009 – 600019	15000	Flu

31-60	Person	600009 – 600019		20000	Gastritis
31-60	Person	600009 – 600019		23000	Flu

Because of this additional constraint on the numerical sensitive attribute, the generalization level of QI attributes increases, which in turn leads to loss of information. The proximity breach is inappropriate for categorical attribute, where different values do not have any sense of proximity.

2.5 Problem Statement

Even though in a QI group G, with L-diverse sensitive categorical values, since, an adversary is restricted to get the actual value of the victim with 1/L% of confidence, there exists a more dangerous situation that, the victim is associated with totally irrelevant information with (1-(1/L))% of confidence. We name this breach as divergence breach.

For example, from the 3- anonymous, 2-diverse patient table 4 released and the voters table 2 released, an adversary may try to find the disease of the victims, Barbie and Girija. Assume that the adversary knows the QI group of both. So, the adversary concludes that the disease of Barbie is Bronchitis or Stomach Cancer. Barbie, being a cancer victim may be happy with this kind of grouping. On the other hand Girija, who is affected with Bronchitis, definitely will not be satisfied with this kind of grouping. Instead, she may prefer to disclose her actual disease for the research purpose. Only because, Girija's QI values are very close to Barbie's QI values, they are placed in one group, to minimize QI generalization loss and to maintain L-Diverse values. This is not a fair grouping, if semantically analyzed.

Also, the proximity problem discussed in [12] gives solution to single numeric sensitive attribute. Our motto is to design a simple micro data sanitization algorithm to preserve privacy against both proximity and divergence attack for any number of sensitive attributes. Also, sanitization should be in such a way that while preserving privacy, maximum information and utility of data is preserved. Also, any type of mining task can be applied, on the transformed table, without even any modification in the mining algorithm.

3. NOTATIONS

Let T be a relation storing private information about a set of individuals. T= $\{t_1, t_2, t_3, \ldots t_n\}$. Each t_i is a tuple of attribute values representing some individual records. Let A= $\{A_1, A_2, \ldots A_m\}$ be a set of attribute in T. $t[A_i]$ represents the value of attribute A_i for tuple t. The set A can be classified into four categories: Identifying Attributes A^i, Sensitive Attributes A^s, Quasi Identifying attributes A^q and Neutral Attributes A^n.

The identifying attributes A^i is not released to the public. The sensitive attribute A^s whose values may be confidential for an individual, may be released for the research purposes, with the individual's concern. A^s is categorized further into numerical sensitive attribute A^{NS} and categorical sensitive attribute A^{CS} based on the data type. Quasi identifier attributes A^q are the set of attributes $\{A^{q1}, A^{q2}, .., A^{qd}\}$ whose values may be published but may reveal personal identity with the aid of external data base. The neutral attributes A^n are neither quasi nor sensitive and therefore can be published as such. Our objective is to publish a table T' derived from T containing all the attributes except A^i and all the tuples in T such a way that T' possesses maximum information.

4. PRIVACY MODEL

The proposed privacy model is a transformed table T', generated from the original table by transforming the sensitive attributes. The data type of the sensitive attributes decides the method of transformation. For generalization of numerical and categorical data, grouping methods and taxonomy based approach were followed in some previous works. Our new definition of privacy breach is, not only leaking the actual sensitive information but also leaking a very short range values of numeric information and totally irrelevant information, even with least probability.

4.1 Graded Grouping

Although various generalization principles such as (k, e) - anonymity [29], Variance Control [15], T-closeness [19] deal with numerical sensitive values A^{NS} to preserve privacy, they failed to protect actual values from proximity attack. Our approach to sensitive numerical attribute is graded grouping as shown in Figure 1.In the figure, the number of grouping (k) is considered as 3 but K may take any value more than 3. Each group is assigned a number name n1, n2, n3….

The proposed approach follows the following steps to convert the actual values into a new form:

1. Fix the number of categories (K) for the given domain and assign a Category Name (CN) 'n' for each. The value of 'n' may be any positive integer value in such a way that n1<n2<n3 and so on.

2. Fix the range (maximum and minimum) for each category $C_1 \ldots C_K$, in such a way that non-overlapping continuous range results. For example, the minimum of second category (Min2) is not equal to the maximum value of first category (Max1) but no actual value exists between these two values.

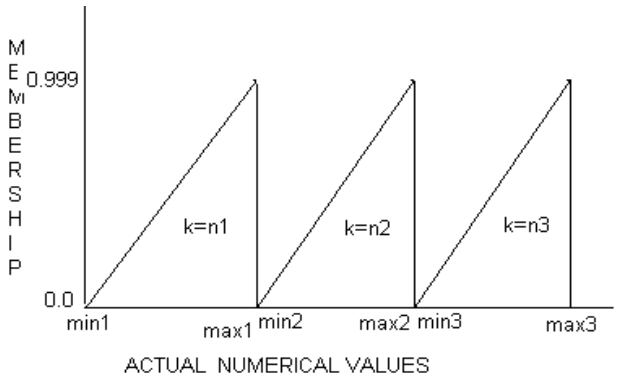
3. Fix the category (C_i), for each actual value, to which it belongs and find the membership value m(x) using

$m(x) = 0.0$ if $x = \min(C_i)$

$=(x-\min(C_i))/(\max(C_i)-\min(C_i))$ if $\min(C_i) > x < \max(C_i)$

$= 0.999$ if $x = \max(C_i)$

4. Replace the actual value with a new value obtained by adding category value (ni) and the membership value m(x).



4.2 Mapping Table

To solve the problem of divergence breach, each categorical sensitive value is assigned a mapping value. The mapping values are assigned in any random order, but mimic the value of categorical attribute. A typical mapping table for the attribute 'Disease' is as shown in table 5. In the table shown Diseases are arranged not in alphabetical order, not in severity order, not in any taxonomy order and assigned mapping values. By this kind of mapping values, the actual values can not be guessed. This mapping table is to be preserved in trusted sever along with the original table T. But, the transformed table T' is released by the server for research purposes.

4.3 Table Transformation Algorithm

Input: Table T with n tuples

Output: Table T' with n tuples with transformed sensitive information

1. All the attributes A_1 to A_m are categorized into four groups A^i, A^q, A^s and A^n (Refer section 3)

2. $T = T - A^i$ \\Suppressing identifying attribute

3. A^s is categorized into A^{NS} and A^{CS} \\ Numerical and Categorical sensitive attribute

4. For i= 1 to n \\ number of records

   $t_i [A^{CS}]$=mapping table value for $t_i [A^{CS}]$

5. Call the function Graded_grup(A^{NS})\\ Refer Section 4.1

5. EXPERIMENTS AND RESULTS

The main goals of the proposed approach are to investigate the performance implication in terms of data mining utility, information content and the privacy level achieved. we have used the adult database from UCI machine learning repository [18] with 32,558 records and the attributes age, education, relation, sex, race, work class, marital status, occupation, country were considered. The attribute age is considered as sensitive numerical attribute and graded grouping is done on this attribute. The attribute education is considered as sensitive categorical attribute and a mapping table is prepared for the same.

We have implemented the table transformation algorithm in Java standard Edition 5.0 and made to run on Intel® Core2 Duo, 1.8 GHz, 1GB RAM system which took only 28sec for generating privacy preserving Adult data set T'. Now T' can be called as sanitized data base.

Table 5. A Sample Mapping Table for Categorical Attribute

ACTUAL DISEASE	MAPPING VALUE
Flu	Illness_1
Stomach Cancer	Illness_2
Bronchitis	Illness_3

Utility of a sanitized database should be measured by how well cross attribute correlations are preserved after sanitization. So users/owners of the sanitized database are interested in workloads that take advantage of attribute correlations within the database, e.g., construction of classifiers.

5.1 Classifier Learning Accuracy

We have used WEKA tool to find the classifier accuracy. Learning accuracies of both the original Adult table and transformed table are exactly the same for the considered class variable. Table 6 lists the classifier learning accuracies (Naïve Bayes Classification) for different class variables. From this table we conclude that data mining utility (classifier accuracy) does not get affected with the proposed approach, which in turn speaks about the truthfulness of data available in T'.

It is critically important to measure both privacy and utility using the same methodology. Otherwise, maximizing utility may lead to privacy violations.

5.2 Rule Generation

The number of rules generated by Adult Data set (T) and transformed table (T') with different class variables are studied, using C-Tree classifier. The confidence and support for rule generation are changed. Results are listed in table 7. From this table we conclude that the number of rules generated remains same, for the given support and confidence. This shows that information is preserved. But, if the sensitive attributes present in the rules, their transformed values are shown. Both the adversary and the researcher *can not guess* the actual values from the rules, which show that privacy is preserved. The researcher is allowed to interpret the results by proving his authentication and results to the central trusted server. A sample set of rules framed by T and T' is shown in figure 2

Table 6. Classifier Accuracy of T and T'

Attributes as class variable	Learning Accuracy (%) of Naïve Bayes Classifier
Work class	73.06
Education	38.34
Marital Status	83.54
Occupation	35.48
Race	85.79

Figure 1. Graded Grouping Transformation for k=3

Country	90.07

Table 7. Comparison of Rule generation

Class Variable	Support %	Confidence %	No. of Rules generated	
			T	T'
Occupation	0	50	4	4
Marital Status	20	50	7	7
Work Class	30	70	13	13
Race	30	80	14	14
Country	30	85	7	7

	ORIGINAL TABLE			TRANSFORMED TABLE	
Rule0	WorkClass = private		Rule0	WorkClass = private	
Rule1	IF	education = 1st-4th	Rule1	IF	hiding_education = general-b
	THEN	WorkClass = private		THEN	WorkClass = private
Rule2	IF	education = assoc-voc	Rule2	IF	hiding_education = general-w
	THEN	WorkClass = private		THEN	WorkClass = private
Rule3	IF	Relation = unmarried	Rule3	IF	Relation = unmarried
	THEN	WorkClass = private		THEN	WorkClass = private
Rule4	IF	age < 42	Rule4	IF	H-age < 1.9
	THEN	WorkClass = private		THEN	WorkClass = private
Rule5	IF	Race = black	Rule5	IF	Race = black
	THEN	WorkClass = private		THEN	WorkClass = private

Figure 2. A sample set of Rules generated with work class as class variable

5.3 Other tasks

We have performed other mining tasks such as clustering and association rule mining on the original and transformed Adult data sets. In Density based clustering, 57% instances are grouped as one cluster, both in T and T'. But, in simple K-means Euclidian distance clustering grouped 63% in one cluster. Since Association Rule mining of WEKA can not handle numeric attribute, the attribute age is omitted from analysis. The number of rules generated by the original and transformed table, using different association rule mining techniques such as apriori, predictive apriori, Tertius is found to be equal, for the considered support and confidence.

6. EVALUATION

In literature, data publishing methods transform the data in some way so that, only a certain types of data mining tasks can be conducted with guaranteed privacy. We do not aim at any specific data mining task. Instead, the proposed approach preserves privacy against any attack (proximity, diversity, linking and homogeneity) and the transformed table is resistive to disclosure of sensitive information, whatever may be the type of mining task. Since, the complexity of the transformation algorithm is linear to the input size, it is applicable to very large dataset. The experiments conducted demonstrate that our method outperforms the existing techniques in terms of execution time while maintaining information and privacy.

6.1 Computation Cost

The overhead of the proposed approach is transformation of the original table T for release. The execution time of our algorithm in producing transformed table T' varies linearly with the size of the table. The algorithm terminates in less than 28 sec for the Adult data set. No other statistical information needs to be maintained. Only the mapping table and category ranges need to be stored safely in the trusted server.

6.2 Quality of Transformed Data

6.2.1 Numeric Data

One way of testing the quality of data is finding the level of matching between the original and transformed data. This provides the nature of relationship between original and transformed data [3]. So, the correlations between numeric sensitive attribute age of original table T and transformed table T' is calculated. It is found that the correlation is above 0.9, whatever may be the number of categories selected, provided all the categories have uniform ranges. However, if different ranges are selected for different categories, the correlation factor gets reduced but still above 0.8. Even if an adversary has some knowledge about the domain, approximate minimum and maximum value of the domain, he will be not be able to infer the actual values. Unless and otherwise knowing the number of categories, numeric name of each category and range of each category no one can guess the actual value. Consider a sample data set derived by graded grouping. The actual value 73 is transformed to 4.8, when the categorical number names (CN) are 1,2,3,4 and 5 but to 8.8 when the CN are 2,4,6,8,10. But in both the cases correlation is maintained.

6.2.2 Categorical Data

Encoding the categorical values will not cause any information loss.To check for the quality of transformed categorical data, regression analysis was done, choosing the transformed and original values as two independent variables. Assuming yet another dependent variable, regression equations are framed for both the cases. The goal of regression analysis is to obtain estimates of the unknown parameters Beta_1, ..., Beta_K which indicate how a change in one of the independent variables affects the values taken by the dependent variable. The coefficient of determination (to know how well the equation fits the data) is found to be equal in the both cases, whatever may be the regression equations framed.

6.3 Information Loss

A variety of information loss metrics have been proposed. But all these metrics were derived based on the equivalence classes generated. Since, the proposed method is not generating any equivalence class; those metrics can not be used to measure information loss. Also, no suppression/ generalization methods applied, the information is preserved 100% in QI attributes. Total information loss in transformed table (IL) is equal to the sum of information loss in numeric sensitive attribute (IL_{NS}) and information loss in categorical sensitive attribute (IL_{CS}). IL_{NS} can be measured using correlation factor between original numerical values and the transformed values.

Table 8. A sample values of graded grouping

Actual Values	Number of Categories K=5	
	C. N.=1,2,3,4,5	C.N.=2,4,6,8,10
15	1	2
25	1.666667	2.666667
30	2	3
37	2.466667	4.466667
40	2.666667	4.666667
58	3.8	6.8
70	4.6	8.6
73	4.8	8.8
90	5.933333	10.93333
Correlation	0.999959	0.997021

For categorical attribute, information loss is measured using height of taxonomy tree to which actual value is generalized. Since, we do the transformation based on mapping values, there is no information loss. Since, there is no grouping or equivalence class; there is no chance for proximity and divergence breach.

6.4 Proximity Breach

In K- anonymization (or improved methods), let *t* be the tuple in T, and G the QI-group in T' that *t* is generalized to. The risk of Proximity Breach of *t*, denoted as $P_b(t)$, equals x / |G| where x is the number of tuples in G whose sensitive values fall in very short interval and |G| the size of G. In our approach, the size of G is equal to the size of table 'n'. Since the denominator value increases, the value of $P_b(t)$, decreases. Even if the transformed values fall in a short range, for a set of records in table T, the actual values can not be guessed, because the transformation depends on number of categories and minimum and maximum values of each category. For example in table 8, the difference in the transformed values corresponding to the actual values 58 and 70 is 0.8 (4.6-3.8), when Category Name (C. N) is 1 to 5. But the difference is 1.8, when the Category Name is 2 to 10. Also, if the number of categories (K) is changed the transformation will take a new value. So, without knowing, K, C. N and range for each category, the actual values can not be guessed and hence proximity breach is avoided, while preserving privacy.

6.5 Divergence Breach

The amount of data distortion occurs by generalization of equivalence class e in K-anonymization is denoted by IL(e)=(|e|*|G|)/|D| where |e| is the number of records in the equivalence class, |D| is the domain size and |G| the amount of generalization. Amount of generalization is zero because of mapping table. Hence the information loss is zero and divergence attack is eliminated.

7. CONCLUSION

Algorithms such as k-anonymity and L-diversity leave all sensitive attributes intact and apply generalization and suppression to the Quasi-identifiers. The goal is to keep the data *truthful* and thus provide good utility for data-mining applications, while achieving less than perfect privacy. But utility is best measured by the success of data mining algorithms such as decision tree learning which take advantage of relationships between attributes.

Also simple anonymization is already widely used in practice. One prime example is clinical trial studies for new drugs in the medical and pharmaceutical domain. Even though the U.S. Food and Drug Administration guidelines are known to be strict, anonymization (or de-identification) is still considered adequate in the clinical trial setting for protecting the privacy of patients participating in the studies. Compared with other transformation techniques, anonymization is simple to carry out, as mapping objects back and forth is easy. Another advantage of anonymization is that it does not perturb data characteristics.

Optimal generalization is NP hard, as well as generalization becomes complex when dimensionality of the table increases [9]. But, the proposed method is extremely efficient because of simplicity in implementation. Since the transformed table preserves the characteristics of the original table, the utility of any mining task is preserved and thereby avoiding need for developing problem specific algoritms.

Being the simple procedure, transformation is done in data owners' site itself and can be supplied for any type mining task. The experiments conducted on the UCI data proved the utility and privacy of data for all typical data mining tasks. Also, the problem of proximity attack and divergence attack is solved by not forming a group or equivalence class.

8. REFERENCES

[1] Adam N. R., Wortmann J. C., "Security-control methods for statistical databases: A comparative study", *ACM Comput. Surv 21(4*), 515–556, 1989

[2] Aggarwal C. C., Yu P. S., "A Condensation approach to privacy preserving data mining", *EDBT Conference*, 2004

[3] Aggarwal C. C., Yu P. S., "On Variable Constraints in Privacy-Preserving Data Mining", *SIAM Conference*, 2005

[4] Agrawal D., Aggarwal C. C., "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms", *ACM PODS Conference*, 2002

[5] Agrawal R., Srikant R.,"Privacy-Preserving Data Mining", *ACM SIGMOD Conference*, 2000

[6] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.,"Disclosure limitation of sensitive rules", *Workshop on Knowledge and Data Engineering Exchange*, 1999

[7] Bayardo. R. J, Rakesh Agrawal, "Data privacy through optimal k- anonymization" , *ICDE*, 217-228,2005

[8] Justin Brickell and Vitaly Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing", *KDD conference*, 2008

[9] C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, 2005

[10] I. Dinur and K. Nissim.,"Revealing information while preserving privacy*", PODS*, pages 202–210, 2003

[11] J. Li, Raymond chi wing wong, Ada Fu, J. pei, "Anonymization by local recoding in data with attribute hierarchical taxonomies", *IEEE transaction on Knowledge and data Engg*, Vol 20, No. 9, pp. 1181-1194,sep 2008

[12] Jiexing Li, Yufei Tao, Xiaokui Xiao, " Preservation of Proximity Privacy in Publishing Numerical Sensitive Data", *ACM SIGMOD*, 2008

[13] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing, *PODS*, 2005

[14] K LeFevre, David J. DeWitt, Raghu Ramakrishnan ,"Incognito: Efficient full domain k – anonymity ", *SIGMOD*, 49-60, 2005

[15] K. Lefevre, D. Dewatt, R. Ramakrishnana, "Workload Aware Anonymization", *ACM KDDM*, 2006

[16] Machanavajjhala A., Gehrke J., Kifer D.,and Venkitasubramaniam M, "l-Diversity: Privacy Beyond k-Anonymity", pp.24-35, *ICDE*, 2006

[17] D. Martin, D.Kifer, A. Machanavajjhala, J. Gehrke, J. Halpern, " Worst-case background knowledge in privacy", *ICDE*, 2007

[18] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases", Available at *www .ics. uci. edu/~ mlearn/MLRepository.html*, University of California, Irvine, 1998

[19] Ninghui Li , Tiancheng Li and Suresh.V, "t-Closeness: Privacy beyond k-anonymity and l-diversity", *ICDE*, 2007

[20] S. R. M. Oliveira and O. R. Zaïane, "Privacy Preservation When Sharing Data For Clustering", *International Workshop on Secure Data Management in a Connected World*, 2004

[21] Benny Pinkas," Cryptographic techniques for privacy preserving data mining", *SIGKDD Explorations*, Vol. 4, Issue.2, pp 12-19, 2002

[22] Pierangela Samarati, "Protecting respondents identities in micro data release", *TKDE*, 13(6), 1010-1027, 2001

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

[24] Wenliang Du, Zhijun Zhang, "A Practical Approach to Solve Secure Multi-party Computation," in *NSPW '02: 2002 workshop on New security paradigms*, pp. 127-135, 2002

[25] Wenliang Du and Zhijun Zhan, "Using Randomized Response Techniques for Privacy-Preserving Data Mining", *SIDKDD* 2003

[26] R.C. W. Wong, J. Li, A. W. C. Fu, K. Wang "(alpha,k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing", *ACM SIGKDD*, pp.139-150,2006

[27] Xiaokui xiao, Yufei Tao, "Personalized Privacy Preservation", *SIGMOD*, 2006

[28] Xiaokui xiao, Yufei Tao, "m-Invariance: Towards Privacy Preserving Re-publication of Dynamic datasets", *SIGMOD*, 2007

[29] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, "Aggregate Query Answering on Anonymized Tables", *ICDE* ,2007