International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 1, January 2013 ISSN 2319 - 4847

Optimizing the Convergence of Data Utility and Privacy in Data Mining

Lambodar Jena[1], Narendra Kumar Kamila[2], Sushruta Mishra[3]

[1]Dept. of CSE. Gandhi Engineering College, Bhubaneswar, India
[2]Dept. of CSE. C.V.Raman college of Engineering, Bhubaneswar, India
[3]Dept. of CSE. Gandhi Engineering College, Bhubaneswar, India

ABSTRACT

Data Mining plays a vital role in today's information world where it has been widely applied in various organizations. The current trend needs to share data for mutual benefit. However, there has been a lot of concern over privacy in the recent years .It has also raised a potential threat of revealing sensitive data of an individual when the data is released publically. Various methods have been proposed to tackle the privacy preservation problem like anonymization, perturbation, generalization and l-diversity. But the natural consequence of privacy preservation is information loss. The loss of specific information about certain individuals may affect the data quality and in extreme case the data may become completely useless. There are methods like cryptography which completely anonymize the dataset and which renders the dataset useless. So the utility of the data is completely lost. We need to protect the private information and preserve the data utility as much as possible. So the objective of this paper is to find an optimum balance between privacy and utility while publishing dataset of any organization. We used K-means algorithm for clustering the dataset and followed by k-anonymization. Privacy preservation is hard requirement that must be satisfied and utility is the measure to be optimized.

Keywords: Privacy preservation, Data mining, K- anonymization, K-means, Utility

1. INDRODUCTION

The amount of data that need to be processed to extract some useful information is increasing. Therefore different data mining methods are adopted to get optimum result with respect to time and utility of data. The amount of personal data that can be collected and analyzed has also increased. Data mining tools are increasingly being used to infer trends and patterns. In many scenarios, access to large amounts of personal data is essential in order for accurate inferences to be drawn. However, publishing of data containing personal information has to be restricted so that individual privacy[1] is not hampered. One possible solution is that instead of releasing the entire database, only a part of it is released which can answer the adequate queries and do not reveal sensitive information. Only those queries are answered which do not reveal sensitive information. Sometimes original data is perturbed and the database owner provides a perturbed answer to each query. These methods require the researchers to formulate their queries without access to any data. Sanitization approach can be used to anonymize[2] the data in order to hide the exact values of the data. But conclusion can't be drawn with surety. Another approach is to suppress some of the data values[5], while releasing the remaining data values exactly. But suppressing the data may hamper the utility. A lot of research work has been done to protect privacy and many models have been proposed to protect databases. Out of them, k-anonymity has received considerable attention from computer scientist. Under k-anonymity, each piece of disclosed data is equivalent to at least k-1 other pieces of disclosed data over a set of attributes that are deemed to be privacy sensitive.

1.1. Methods of Data Mining

The amount of data that need to be processed to extract some meaningful information is increasing. So the methods used for extracting information from huge amount of data must be optimum. The various data mining algorithms can be classified into several categories[1].

1.1.1. Additive-Noise-based Perturbation Techniques

Random noise is added to the actual data in additive-noise-based perturbation technique. The privacy is measured by evaluating how closely the original values of a modified attribute can be determined. In particular, if the perturbed value of an attribute can be estimated, with a confidence c, to belong to an interval [a, b], then the privacy is estimated by (b−a) with confidence c. However, this metric does not work well because it does not take into account the distribution of the original data along with the perturbed data.

1.1.2. Multiplicative-Noise-based Perturbation Techniques

Additive random noise can be filtered out using certain signal processing techniques[2] with very high accuracy. This problem can be avoided by using random projection-based multiplicative perturbation techniques[3]. Instead of adding some random values to the actual data, random matrices are used to project the set of original data points to a randomly chosen lower-dimensional space. However, the transformed data still preserves much statistical aggregate regarding the original dataset so that certain data mining tasks can be performed on the transformed data in a distributed environment (data are either vertically partitioned or horizontally partitioned) with small errors. High degree of privacy of original data is ensured in this approach. Even if the random matrix is disclosed, it only approximate value of original data can be estimated. It is impossible to get back the original data. The variance of the approximated data is used as privacy measure.

1.1.3. k- Anonymization Techniques

A database is k-anonymous with respect to quasi-identifier attributes (defined later in this thesis) if there exist at least k transactions in the database having the same values according to the quasi-identifier attributes[4, 5]. In practice, in order to protect sensitive dataset T, before releasing T to the public, T is converted into a new dataset T* that guarantees the k-anonymity property for a sensible attribute. This is done by generalizations and suppression on quasi-identifier attributes. Therefore, the degree of uncertainty of the sensitive attribute is at least 1/k.

1.1.4. Statistical-Disclosure-Control-based Techniques

To anonymize the data to be released (such as person, household and business) which can be used to identify an individual, additional information publicly available need to be considered[6]. Among these methods specifically designed for continuous data, the following masking techniques are described: additive noise, data distortion by probability distribution, resampling, rank swapping, etc. The privacy level of such method is assessed by using the disclosure risk, that is, the risk that a piece of information be linked to a specific individual.

1.1.5. Cryptography-based Techniques

The cryptography-based technique usually guarantees very high level of data privacy[14]. Generally solution is based on the assumption that each party first encrypts its own item sets using commutative encryption, then the already encrypted item sets of every other party. The two communicating party must share a common key which is used for encryption and decryption. Sometimes two key is used known as public key and private key. Public key is known to everybody that wants to communicate with you and private key is used for decryption in a secure communication. Though cryptography-based techniques can well protect data privacy, they may not be considered good with respect to other metrics like efficiency.

1.2. Privacy

Privacy means how an individual controls his personal information from access of others. From another point of view, privacy may be how the data is collected, shared and used by the customers. So definition of privacy varies from one environment to the other. So the definition of privacy [1] is:

- Privacy as the right of a person to determine which personal information about himself/ herself may be communicated to others.
- Privacy as the control over access to information about oneself.
- Privacy as limited access to a person and to all the features related to the person.

1.3. Data Utility

The utility of the data must be preserved to certain extent at the end of the privacy preserving process, because in order for sensitive information to be hidden, the database is essentially modified through the changing of information (through generalization and suppression) or through the blocking of data values. Sampling is a privacy preserving technique which does not modify the information stored in the database, but still, the utility of the data falls, since the information is not complete in this case. As we go on changing the data for preserving privacy, the less the database reflects the domain of interest. So, one of the evaluation parameter for the measuring data utility should be the amount of information that is lost after the application of privacy preserving process. Of course, the measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed. Information loss in the context of association rule mining[7] will be measured either in terms of the number of rules that were both remaining and lost in the database after sanitization, or even in terms on the reduction/increase in the support and confidence of all the rules. For the case of classification, we can use metrics similar to those used for association rules. Finally, for clustering, the variance of the distances among the clustered items in the original database and the sanitized database can be the basis for evaluating information loss in this case.

2. Generalization and Suppression

Various method have been proposed for providing anonymity in the release of micro data, the k-anonymity proposal focuses on two techniques in particular: generalization and suppression[5], which, unlike other existing techniques, such as scrambling or swapping, preserve the truthfulness of the information. In the following paragraph we have described it in detail.
The mapping is stated by means of a generalization relationship ≤d. Given two domains Di and Dj Є Dom, Di ≤d Dj states that values in domain Dj are generalizations of values in Di. The generalization relationship ≤d defines a partial order on the set Dom of domains, and is required to satisfy the following conditions [4, 6]

C1: $\forall$ Di, Dj, Dz ∊ Dom:
(Di ≤D Dj), (Di ≤D Dz) => (Dj ≤D Dz) V (Dz ≤D Dj) ,

C2: all maximal elements of Dom are singleton.

Condition C1 states that for each domain Di, the set of domains generalization of Di is totally ordered and, therefore, each Di has at most one direct generalization domain Dj. It ensures determinism in the generalization process. Condition C2 ensures that all values in each domain can always be generalized to a single value. The definition of a generalization relationship implies the existence, for each domain D Dom, of a totally ordered hierarchy, called domain generalization hierarchy, denoted DGHD. A value generalization relationship is denoted as ≤v which associates with each value in domain Di a unique value in domain Dj, direct generalization of Di. The value generalization relationship implies the existence, for each domain D, of a value generalization hierarchy, denoted VGHD.

2.1. k-Minimal Generalization (with Suppression)

Definition 3 (Generalized table - with suppression). Let Ti and Tj be two tables defined on the same set of attributes. Table Tj is said to be a generalization (with tuple suppression) of table Ti, denoted Ti ≤ Tj, if:
1. | Tj| ≤ | Ti|
2. The domain dom(A, Tj) of each attribute A in Ti is equal to, or a generalization of, the domain dom(A, Ti) of attribute A in Ti
3. It is possible to define an injective function associating each tuple t_j in Tj with a tuple t_i in Ti, such that the value of each attribute in t_i is equal to, or a generalization of, the value of the corresponding attribute in t_i.

2.2. k-Anonymity and k-Anonymous Tables

The concept of k-anonymity requires that the released private table (PT) should be indistinguishably related to no less than a certain number of respondents which is followed by all statistical community and by agencies. The set of attributes included in the private table, also externally available and therefore exploitable for linking, is called quasi-identifier. The k-anonymity requirement [6] states that every tuple released cannot be related to fewer than k respondents.
Definition 1 (k-anonymity requirement): Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.
To guarantee the k-anonymity requirement, k-anonymity requires each quasi identifier value in the released table to have at least k occurrences [6] .
Definition 2 (k-anonymity): Let T(A1………Am) be a table, and QI be a quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI if each sequence of values in T[QI] appears at least with k occurrences in T[QI].

2.3. Privacy Principles

The information published in the anonymized table is prone to attack due to the background knowledge of the adversary [9]. So the private information might be revealed in two ways: Positive disclosure and Negative disclosure.

2.3.1. Positive disclosure

The original table T published after anonymization as T* results in a positive disclosure if the adversary can correctly identify the value of a sensitive attribute with high probability; i.e., given a $\delta > 0$, there is a positive disclosure if β (q, s, T*) > $(1 - \delta)$ and there exists t Є T such that t[Q] = q and t[S] = s.

2.3.2. Negative disclosure

The original table T after anonymization is published as T* results in a negative disclosure if the adversary can correctly eliminate some possible values of the sensitive attribute with high probability; i.e., given an Є > 0, there is a negative disclosure if β (q, s, T*) < Є and there exists a t Є T such that t[Q] = q but t[S] != s.

All positive disclosures are not disastrous neither all negative disclosure[9]. If the prior belief was that α (q, s) > 1−δ, the adversary would not have learned anything new. Hence, the ideal definition of privacy can be based on the following principle:

2.3.3. Uninformative Principle

The published table should provide the adversary with little additional information beyond the background knowledge. In other words, there should not be a large difference between the prior and posterior beliefs.
Suppose the published table T* has two constants $\rho 1$ and $\rho 2$, we say that a ($\rho 1$, $\rho 2$)- privacy breach has occurred when either α (q, s) < $\rho 1$ □ β (q, s, T*) > $\rho 2$ or when α (q, s) > 1 − $\rho 1$□ β (q, s, T*) < 1−$\rho 2$. If a ($\rho 1$, $\rho 2$) privacy breach has not occurred, then table T* satisfies ($\rho 1$, $\rho 2$)-privacy.

2.4. l-Diversity

The l-diversity model is a very useful model for preventing attribute disclosure [9].
l-Diversity Principle: A q*-block is l-diverse if it contains at least l well-represented values for the sensitive attribute S. A table is l-diverse if every q*-block is l-diverse. The l-diversity principle advocates ensuring l well-represented values for the sensitive attribute in every q-block, but does not clearly state what well-represented means. It has the following properties:

- Knowledge of the full distribution of the sensitive and non-sensitive attributes is not required in l-diversity.
- l-diversity does not even require the data publisher to have as much information as the adversary. The larger the value of l, the more information is needed to rule out possible values of the sensitive attribute.
- Different adversaries can have different background knowledge leading to different inferences. It simultaneously protects against all of them without the need for checking which inferences can be made with which levels of background knowledge.

2.4.1. Distinct l-diversity:

The term "well represented" in the definition of l-diversity would be to ensure there are at least l distinct values for the sensitive attribute in each equivalence class. Distinct l-diversity does not prevent probabilistic inference attacks. It may happen that in an anonymized block one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value.

2.4.2. Entropy l-diversity:

The entropy of an equivalence class E is defined to be

$$(E) = -\sum (E,s) \log p(E,s) \text{ , for } s\epsilon S$$

Where S is the domain of the sensitive attribute, and p(E, s) is the fraction of records in E that have sensitive value s. A table is said to have entropy l-diversity if for every equivalence class E, Entropy(E) ≥ log l. Entropy l- diversity is strong than distinct l-diversity. In order to have entropy l-diversity for each equivalence class, the entropy of the entire table must be at least log (l). Sometimes this may too restrictive, as the entropy of the entire table may be low if a few values are very common.

2.4.3. Recursive (c, l)-diversity:

Recursive (c, l)-diversity ensure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Let m be the number of values in an equivalence class, and r_i, $1 \leq i \leq m$ be the number of times that the i^{th} most frequent sensitive value appears in an equivalence class E. Then E is said to have recursive (c, l)-diversity if $r_1 < c(r_l + r_{l+1} + ... + r_m)$. A table is said to have recursive (c, l)-diversity if all of its equivalence classes have recursive (c, l)-diversity.

2.5. Attacks on l-diverse data

Skewness Attack:
l-diversity does not prevent attribute disclosure if the overall distribution is skewed. Consider an equivalence class has an equal number of positive records and negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive (c, 2)-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population. Now consider an equivalence class that has 49 positive records and only 1 negative record. It would be distinct 2-diverse and has higher entropy than the overall table (and thus satisfies any Entropy l-diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 positive and 49 negative record, even though the two classes present very different levels of privacy risks.

Similarity Attack:

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example. Table 2.2 is the original table, and Table 2.3 shows an anonymized version satisfying distinct and entropy 3-diversity. There are two sensitive attributes: Salary and Disease. Suppose one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's salary is in the range [3K–5K] and can infer that Bob's salary is relatively low. This attack applies not only to numeric attributes like "Salary", but also to categorical attributes like "Disease". Knowing that Bob's record belongs to the first equivalence class enables one to conclude that Bob has some stomach-related problems, because all three diseases in the class are stomach-related. This leakage of sensitive information occurs because while l-diversity requirement ensures "diversity" of sensitive values in each group, it does not take into account the semantical closeness of these values.

Table 2.2: Original Salary/Disease table

	ZIP code	Age	Salary	Disease
	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 2.3: A 3-diverse version of Table 2.2

	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

3. ALGORITHM

The proposed algorithm uses generalization and tuple suppression over quasi-identifiers to obtain a k-anonymized table with maximum suppression of MaxSup tuples. This algorithm uses binary search on the generalization hierarchy to save time. It assumes that a table PT with more than k attributes is present which is to be k-anonymized. In this approach concept of distance vector is induced and exploited. Let PT be a table and x,y Є PT be two tuples such that x =(v1,......,vn) and y=(v1,…,vn) where vi and vi are values in domain Di The distance vector between x and y is the vector Vx,y = [d1……..dn] where di is the (equal) length of the two paths from vi and vi to their closest common ancestor in the value generalization hierarchy VGHDi (or, in other words, the distance from the domain of vi and vi to the domain at which they generalize to the same value vi).

The one-pass k-means algorithm(OKA) is derived from the standard k-means algorithm[12] but it runs for one iteration. This algorithm has two stages first is the clustering stage and second is the adjustment stage.

Clustering stage:

Let n be the total number of records present in the table T to be anonymized. Then $N = \lfloor n/k \rfloor$, where k is the value of k-anonymity. Clustering stage proceeds by sorting all the records and then randomly picking N records as seeds to build clusters. Then for each record r remaining in the dataset, algorithm checks to find the cluster of which this record is closest and assigns the record to the cluster and updates its centroid. The difference between the traditional k-means algorithm and OKA is that in OKA whenever a record is added to the cluster its centroid is updated thus improving the assignments in future and the centroid represents the real centre of the cluster. In OKA the records are first sorted according to the quasi-identifiers thus making sure that similar tuples are assigned to the same cluster. The algorithm has a complexity of O (n^2k) .

Adjustment Stage:
In the clustering stage the clusters that are formed can contain more than k tuples and there can be some clusters containing less than k tuples, therefore when these clusters are anonymized will not satisfy condition for k-anonymity. These clusters need to be resized to contain at least k tuples. The goal of this adjustment stage is to make the clusters contain at least k records, while minimizing the information loss. This algorithm first removes the extra tuples from the clusters and then assigns those tuples to the clusters having less than k tuples. The removed tuples are farthest from the centroid of the cluster and while assigning the tuples to the clusters it checks the cluster which is closest to the tuple before assigning it, thus minimizing the information loss. If no cluster contains less than k tuples and some records are left they are assigned to this respective closest clusters. The time complexity of this algorithm is O(n^2k) .

Algorithm: Clustering stage	**Algorithm: Adjustment Stage**
Input: a set T of n records; the value k for k-anonymity Output: a partitioning P = {$P1$, . . . , PK} of T 1. Sort all records in dataset T by their quasi-identifiers; 2. Let N := $\lfloor n/k \rfloor$; 3. Randomly select N distinct records $r1$, . . . , rN belongs to T ; 4. Let $P\ i$:= {ri} for i = 1 to N; 5. Let T := T \ {$r1$, . . . ,rN }; 6. While (T != null ;) do 7. Let r be the first record in T ; 8. Calculate the distance between r to each Pi; 9. Add r to its closest Pi; update centroid of Pi; 10. Let T := T \ {r}; 11. End of While	Input: a partitioning P = {$P1$, . . . , PK } of T Output: an adjusted partitioning P = {$P1$, . . . , PK } of T 1. Let R := null ; 2. For each cluster P belongs to p with |P| > k do 3. Sort tuples in P by distance to centroid of P; 4. While (|P| > k) do 5. r belongs to P is the tuple farthest from centroid of P; 6. Let P := P \ {r}; R := R [{r}; 7. End of While 8. End of For 9. While (R != null) do 10. Randomly select a record r from R; 11. Let R := R \ {r}; 12. If P contains cluster Pi such that |Pi| < k then 13. Add r to its closest cluster Pi satisfying |Pi| < k; 14. Else 15. Add r to its closest cluster; 16. End If 17. End of While

4. IMPLEMENTATION AND RESULTS

For implementation of the work we have used the following tools.
NetBeans: NetBeans is an integrated developing environment(IDE) written in the Java programming language, which can be used for developing with java, JavaScript, PHP, Python, Ruby, Groovy, C, C++ and much more. We have used NetBeans 6.0 to implement the algorithms as described in the previous chapter using java.
WEKA: Waikato Environment for Knowledge Analysis (WEKA) is a popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License. It contains a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. We have used WEKA 3.6 for clustering and classification.

4.1. Experimental Set-up:

We carried out the experiments on the standard adult database from **UCI (University of California Irvine)** machine learning repository with 32,564 records. It contains numerical as well as categorical attributes which is suitable for generalization required in our experiment. It contains the following 15 attributes such as age, work class, fnlwgt, education, education num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native country and income, which may take various values .The algorithms were implemented in java and executed on a workstation with Intel Dual Core Processor, 1.80 GHz and 1.00 GB of RAM on Window XP SP2 platform.
Clustering: Clustering of the database is done using WEKA. We have used K-means clustering for our experiment. The clustered results produced by WEKA are saved for further use in the experiment. Figure 4.3 shows the clustering results produced by WEKA.

International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 1, January 2013 ISSN 2319 - 4847

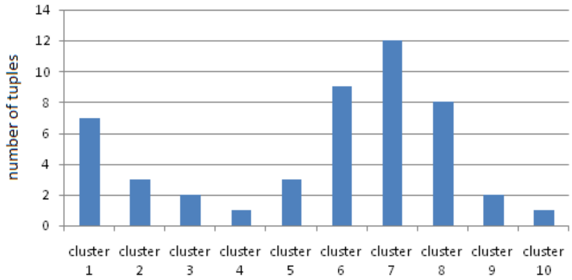

Figure 4.3: Clusters generated by WEKA

Figure 4.3 shows that the clusters are not uniform and cannot be used for k-anonymization. Thus we need to adjust the size of these clusters so that each cluster contains at least k tuples.

4.2. Generalization:

Generalization is done on the clustered dataset from the K-means algorithm. Out of the total 15 attributes we considered 5 attributes as quasi-identifiers and rest as sensitive attributes.

Generalization rules: For age which is a numerical attribute mean of all the tuple values is taken.

$$\text{Mean age} = \frac{\sum_{i=1}^{k} t(i)}{k}$$

Generalization hierarchy is done for education, native-country, race and work class. These generalization hierarchies are used for k-anonymizing evenly clustered data.

4.3. Methodology Used for determining Utility and Privacy:

Utility: To determine the utility of the dataset we have used Decision stump algorithm for classification which is already implemented in WEKA. Decision stump is a machine learning model consisting of a single-level decision tree with a categorical or numeric class label. The results produced by WEKA clearly show percentage of tuples that can be correctly classified using the algorithm.

Privacy: To determine the extent of privacy preserved by the dataset we counted the number of attributes whose values are completely suppressed. The generalization suppresses the attribute value by generalizing the whole domain to "xxx". Queries used for calculating privacy for our generalization hierarchy are as follows:

- Select count (*) from dataset where education = 'xxx' ;
- Select count (*) from dataset where country = 'xxx' ;
- Select count (*) from dataset where race ='xxx' ;
- Select count (*) from dataset where work class = 'xxx' ;

Percentage of privacy preserved in the anonymized dataset is given by the formula:

$$\text{Privacy \%} = \left(\frac{\text{Total number of suppressed values}}{\text{Total number of quasi-identifier values}}\right)*100$$

Experiment 1: Anonymizing sample dataset containing 1000 tuples

In the first experiment we considered only six attributes, age, education, marital status, occupation, race and native-country for our analysis. We randomly selected 1000 tuples from the dataset for anonymization to determine how utility varies with privacy. Age, education, race and country are considered as quasi-identifiers and other two as sensitive attributes. First we used WEKA to arrange the data into clusters according to the value of k. As described earlier the clusters produced by WEKA may contain less than k tuples, thus an adjustment is required so that each cluster contains at least k tuples. Before applying the generalization clusters are adjusted so that each cluster contains at least k tuples. After adjusting the clusters, k-anonymization is done based on the generalization hierarchy. We have implemented k anonymization algorithm based on OKA to generalize the adjusted clusters. Figure 4.4 shows a 5-k anonymized dataset obtained after anonymizing.

Results: For evaluating utility, we performed the classification mining on the k-anonymized dataset (DT). Classification was performed by using WEKA Data Mining Software considering native-country as classification variable. We considered the percentage of correctly classified tuples as the utility of the dataset. Table 4.9 shows the results produced by the WEKA on using decision stump algorithm for a 3-anonymized dataset. Privacy was calculated by counting the number of tuples which are generalized to xxx. Privacy percentage is calculated as described in section 4.4. Privacy and utility was calculated by varying the value of k. The balancing point between utility and privacy is the point where privacy and utility curves intersect or tend to converge. Figure 4.5 shows the variation of utility and privacy with k. It clearly follows from the figure that on increasing the value of k privacy provided by the dataset increases but utility decreases. For this sample dataset the balancing point comes between k=9 and k=10, and utility of the dataset at balancing point is around 60%.

```
28,Graduate,Divorced,Adm-clerical,xxx,North_America,cluster1
28,Graduate,Married-civ-spouse,Sales,xxx,North_America,cluster1
28,Graduate,Never-married,Other-service,xxx,North_America,cluster1
28,Graduate,Married-civ-spouse,?,xxx,North_America,cluster1
28,Graduate,Married-civ-spouse,Sales,xxx,North_America,cluster1
34,xxx,Never-married,Craft-repair,xxx,North_America,cluster2
34,xxx,Married-civ-spouse,Exec-managerial,xxx,North_America,cluster2
34,xxx,Divorced,Other-service,xxx,North_America,cluster2
34,xxx,Never-married,Craft-repair,xxx,North_America,cluster2
34,xxx,Separated,Other-service,xxx,North_America,cluster2
48,xxx,Never-married,Craft-repair,xxx,North_America,cluster3
48,xxx,Divorced,Transport-moving,xxx,North_America,cluster3
48,xxx,Married-spouse-absent,Craft-repair,xxx,North_America,cluster3
48,xxx,Widowed,Other-service,xxx,North_America,cluster3
48,xxx,Married-civ-spouse,Transport-moving,xxx,North_America,cluster3
29,Graduate,Widowed,Other-service,xxx,North_America,cluster4
29,Graduate,Never-married,Sales,xxx,North_America,cluster4
29,Graduate,Never-married,Handlers-cleaners,xxx,North_America,cluster4
29,Graduate,Never-married,Handlers-cleaners,xxx,North_America,cluster4
29,Graduate,Never-married,Other-service,xxx,North_America,cluster4
21,Graduate,Never-married,?,xxx,North_America,cluster5
21,Graduate,Never-married,Sales,xxx,North_America,cluster5
21,Graduate,Never-married,Other-service,xxx,North_America,cluster5
```

Table 4.9: WEKA Classification Result for 3-Anonymized Dataset

Correctly Classified Instances	864	84.6847 %
Incorrectly Classified Instances	153	15.3153 %

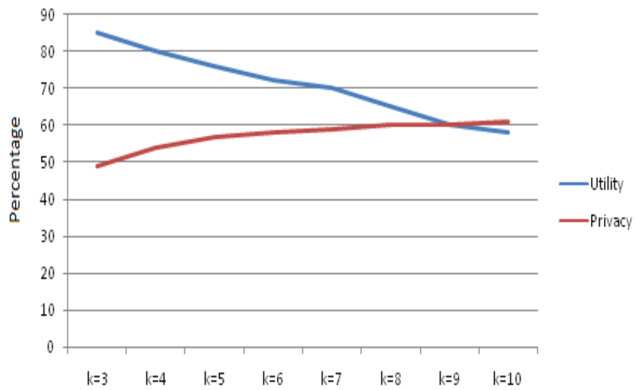

Figure 4.5: Variation of Utility And privacy with anonymization(1000 tuples)

Experiment 2:

In the second experiment we considered all the attributes for our analysis, to study the effect of more number of attributes on the privacy and the utility of the k-anonymized dataset. We randomly selected 1000 tuples from the dataset for anonymization to determine how utility varies with privacy. Age, work class, education, race and native-country are considered as quasi-identifiers and all other attributes as sensitive attributes. Similar steps were followed as in experiment 1 to study the variation of utility and privacy on varying k value. Figure 4.6 shows a 5-k anonymized dataset obtained after anonymizing.

```
20,Private,257509,HS-grad,9,Never-married,Craft-repair,Own-child,White,Male,0,0,40,United-States,<=50K.,cluster1
20,Private,479296,HS-grad,9,Never-married,Handlers-cleaners,Own-child,White,Male,0,0,40,United-States,<=50K.,cluster1
20,Private,169699,HS-grad,9,Never-married,Adm-clerical,Not-in-family,White,Female,0,0,40,United-States,<=50K.,cluster1
20,qqq,30796,HS-grad,9,Never-married,qqq,Own-child,White,Female,0,0,20,United-States,<=50K.,cluster2
20,qqq,334105,HS-grad,9,Never-married,qqq,Not-in-family,White,Female,0,0,40,United-States,<=50K.,cluster2
20,qqq,273989,HS-grad,9,Never-married,Transport-moving,Own-child,White,Male,0,0,40,United-States,<=50K.,cluster2
20,qqq,38455,HS-grad,9,Never-married,qqq,Unmarried,White,Male,0,0,40,United-States,<=50K.,cluster3
20,qqq,419984,HS-grad,9,Divorced,Other-service,Unmarried,White,Female,0,0,25,United-States,<=50K.,cluster3
20,qqq,191948,HS-grad,9,Married-civ-spouse,Other-service,Other-relative,White,Female,0,0,40,United-States,<=50K.,cluster3
19,Private,123007,HS-grad,9,Never-married,Adm-clerical,Other-relative,White,Male,0,1901,30,North_America,<=50K.,cluster4
19,Private,237956,HS-grad,9,Never-married,Protective-serv,Own-child,White,Male,0,0,40,North_America,<=50K.,cluster4
19,Private,179020,HS-grad,9,Never-married,Machine-op-inspct,Own-child,White,Female,0,0,48,North_America,<=50K.,cluster4
18,Private,366154,HS-grad,9,Never-married,Other-service,Not-in-family,White,Male,0,0,30,United-States,<=50K.,cluster5
18,Private,41879,HS-grad,9,Never-married,Handlers-cleaners,Unmarried,White,Male,0,0,25,United-States,<=50K.,cluster5
18,Private,228216,HS-grad,9,Never-married,Handlers-cleaners,Own-child,White,Male,0,0,20,United-States,<=50K.,cluster5
18,qqq,217439,HS-grad,9,Never-married,Other-service,Not-in-family,White,Female,0,0,28,United-States,<=50K.,cluster6
18,qqq,170183,HS-grad,9,Never-married,Adm-clerical,Own-child,White,Female,0,0,10,United-States,<=50K.,cluster6
18,qqq,240183,HS-grad,9,Never-married,qqq,Own-child,White,Female,0,0,45,United-States,<=50K.,cluster6
17,qqq,40299,11th,7,Never-married,Sales,Own-child,White,Female,0,0,25,United-States,<=50K.,cluster7
17,qqq,61838,11th,7,Never-married,Farming-fishing,Own-child,White,Male,0,0,40,United-States,<=50K.,cluster7
17,qqq,143331,11th,7,Never-married,qqq,Own-child,White,Male,0,0,40,United-States,<=50K.,cluster7
27,Private_emp,28544,yyy,7,Never-married,Sales,Not-in-family,White,Female,0,0,20,United-States,<=50K.,cluster8
27,Private_emp,190968,yyy,4,Never-married,Craft-repair,Own-child,White,Male,0,0,27,United-States,<=50K.,cluster8
27,Private_emp,98806,yyy,4,Married-civ-spouse,Craft-repair,Husband,White,Male,0,0,38,United-States,<=50K.,cluster8
30,Private,23778,Some-college,10,Never-married,Exec-managerial,Not-in-family,White,Male,4416,0,40,United-States,<=50K.,cluster9
30,Private,97306,Some-college,10,Never-married,Sales,Not-in-family,White,Female,0,0,40,United-States,<=50K.,cluster9
30,Private,143766,Some-college,10,Never-married,Craft-repair,Not-in-family,White,Male,0,0,40,United-States,<=50K.,cluster9
```

Figure 4.6: 3-anonymized Dataset

Results: As described in previous experiment privacy and utility were calculated by varying the value of k. The balancing point between utility and privacy is the point where privacy and utility curves intersect or tend to converge. Figure 4.7 shows the variation of utility and privacy with k. For this sample dataset the balancing point comes between k=11 and k=12, and utility of the dataset at balancing point is around 52%. Thus on increasing the number of quasi-identifiers considered for analysis the balancing point is shifts down and values of k at which balance is achieved increases.

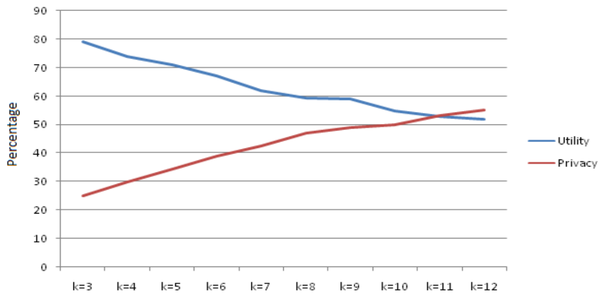

Figure 4.7: Variation of Utility And privacy with anonymization(1000 tuples)

Anonymizing sample dataset containing 3000 tuples:
In this experiment we took 3000 tuples from the adult dataset and carried out the same experiment. We considered all the attributes for our analysis, to study the effect of more number of tuples on the privacy and the utility of the k-anonymized dataset. Age, work class, education, race and native-country are considered as quasi-identifiers and all other attributes as sensitive attributes. Similar steps were followed as in experiment 1 to study the variation of utility and privacy on varying k value. Figure 4.8 shows variation of utility and privacy on varying value of k. For this sample dataset the balancing point comes between k=10 and k=11, and utility of the dataset at balancing point is around 50%.

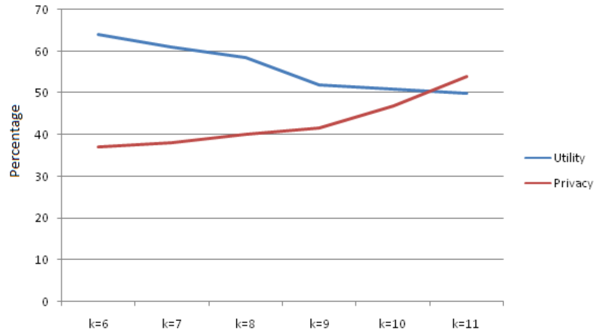

Figure 4.8: Variation of Utility and Privacy with anonymization (3000 tuples)

5. CONCLUSION

In order to improve the privacy offered by the dataset, utility of the data suffers. On conducting the experiments we found that the balancing point between utility and privacy depends on the dataset and value of k cannot be generalized for all datasets such that utility and privacy are balanced. On varying the number of sensitive attributes in a dataset the balancing point varies. We found that if number of quasi-identifiers increases balancing point moves down and balance between utility and privacy occurs at a higher value of k. Thus if a dataset contains more number of quasi-identifiers then the utility as well as privacy attained at balancing point will be less than the dataset having fewer quasi-identifiers. We also studied the effect of number of tuples in the dataset on the balancing point and found that as the number of tuple increases there is slight shift in the balancing point and the value of k for which balancing occurs. Thus we can approximately predict the balancing point for a huge dataset by conducting experiment on a sample dataset. Other privacy preserving algorithms can also be used to find a balancing point between privacy and utility.

REFERENCES

[1] E. Bertino, D. Lin, W. Jiang . A Survey of Quantification of Privacy. In: Privacy-Preserving Data Mining. Springer US, Vol 34, pp. 183-205,2008.
[2] R. J. Bayardo, R. Agrawal. Data privacy through optimal k-anonymization. In: Proc. of the 21st International Conference on Data Engineering, IEEE Computer Society, pp. 217-228, 2005.
[3] K. Liu, H. Kargupta, J. Ryan .Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering, Vol 18(1), pp. 92–106, 2006.
[4] P. Samarati. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, VOl 13(6), pp. 1010–1027, 2001.
[5] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol 10(5), pp. 571–588, 2002.
[6] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In: Secure Data Management in Decentralized Systems. Springer US, Vol 33, pp. 323-353, 2007.
[7] V. S. Verykios, E. Bertino, I. N. F. L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in Privacy Preserving Data Mining. ACM SIGMOD Record, Vol 33(1), pp. 50-57, 2004.
[8] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol 10 (5), pp. 557-570, 2002.
[9] A. Machanavajjhala, J. Gehrke ,D. Kifer, M. Venkitasubramaniam . ℓ-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data, Vol 1(1), Article: 3, 2007.
[10] R. Agrawal, R. Srikant . Privacy preserving data mining. ACM SIGMOD Record, Vol 29(2), pp. 439–450, 2000.
[11] M. R. Z. Mirakabad and A. Jantan . Diversity versus Anonymity for Privacy Preservation. The 3rd International Symposium on Information Technology (ITSim2008), Vol 3, pp. 1-7, 2008.
[12] J.Lin, and M. Wei. An Efficient Clustering Method for k-Anonymization. In: Proceedings of the international workshop on Privacy and anonymity in information society, Vol. 331, pp. 46-50, 2008.
[13] UCI Repository of machine learning databases, University of California, Irvine. http:/archive.ics.uci.edu/ml/.
[14] Lambodar Jena, Narendra K. Kamila, Ramakrushna Swain, Tanmay ku. Das;"Secure Privacy Preserving Association Rule Data Mining: A Cryptographic Approach" ; In the International Journal of Computer Application; Volume 2 , Issue 1,Pg.1-14, 2011.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Vol 11(1), 2009.
[16] Nissim Matatov, Lior Rokach, Oded Maimon. Privacy-preserving data mining: A feature set partitioning approach. Information Sciences 180 (2010) 2696–2720
[17] M. Meints, J. Moller, Privacy preserving data mining – a process centric view from a European perspective, Available online at http://www.fidis.net, 2004.
[18] Shibnath Mukherjee, Madhushri Banerjee, Zhiyuan Chen *, Aryya Gangopadhyay. A privacy preserving technique for distance-based classification with worst case privacy guarantees. Data & Knowledge Engineering 66 (2008) 264–288

AUTHOR

Lambodar Jena received his M.Tech degree in Computer Science and Application from Indian School of Mines(ISM), Dhanbad in the year 2000 and currently pursuing doctorate(Ph.D) degree from Utkal University, India . His research interest includes data mining ,data privacy, image mining,web mining and wireless sensor networking. He has several publications in national/international journals and conference proceedings.

Narendra Kumar Kamila received his M.Tech degree in Computer Science and Engineering from Indian Institute of Technology, Kharagpur and doctorate(Ph.D) degree from Utkal University, India in the year 2000. Later he had visited USA for his post doctoral work at University of Arkansas in 2005. His research interest includes artificial intelligence, data privacy, image processing and wireless sensor networking. He has several publications in journal and conference proceedings. His professional activities include teaching computer science, besides he organizes many conferences, workshops and faculty development programs funded by All India Council for Technical Education, Govt. of India. Dr. Kamila has been rendering his best services as editorial board member and editor-in-chief of many international journal and conference proceedings.

Sushruta Mishra received his B.Tech in Computer Science and Engineering from ITER, Bhubaneswar,India in the year 2009 and M.Tech in Computer Science and Engineering from IIIT, Bhubaneswar,India in the year 2012. His research interest includes data mining and wireless sensor networking. He has several publications in national/international journals and conference proceedings.