Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining

Praveena Priyadarsini
Dept of CSE.
Avinashilingam University
Coimbatore – 641 108

M.L.Valarmathi, PhD.
Dept of CSE.
Govt College of Technology
Coimbatore – 641 028

S.Sivakumari, PhD.
Dept of CSE.
Avinashilingam University
Coimbatore – 641 108

ABSTRACT

Privacy-preserving in data mining refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure and hence protecting individual data records and their privacy. Data perturbation is a privacy preservation technique which does addition / multiplication of noise to the original data. It performs anonymization based on the data type of sensitive data. Generalization is a technique were quasi identifiers data are replaced by some other more general term. In this paper privacy protection is applied to high dimensional datasets like Adult and Census. For ranking the attributes, information gain feature subset selection method is used. The high ranking attributes with sensitive information are set as quasi identifiers of the datasets. A hybrid perturbation technique is used to perturb categorical and numeric attributes of both the datasets and the utility of the datasets is measured using accuracy on data mining functionalities. The data distortion is measured using maintenance of Rank of Features (CK) between the original and perturb datasets. Experimental results show that utility of the perturbed datasets comparable with the original dataset and the Census dataset has comparable CK value than adult dataset.

General Terms

Information privacy, Data mining, Security

Keywords

Data mining, Privacy preservation, perturbation, generalization, utility, classifications, clustering, maintenance of Rank of Features

1. INTRODUCTION

Data mining uses large database with sensitive information and reveals useful patterns from it. The patterns may compromise confidentiality and privacy obligations of an individual or a organization. More the advanced data mining techniques used more the risk of exposing private data. Thus providing security to sensitive data against unauthorized access has been an important area of research in data mining. The concept of privacy violation in data mining concentrates on the validity of result for a private input data [1]. Privacy-preserving data mining (PPDM) is the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Privacy preservation is primarily concerned with protecting against disclosure of individual data records. Most traditional data mining techniques analyze and model the data set statistically, in aggregation; while privacy preservation is primarily concerned with protecting against disclosure individual data records [2]. There are various privacy preservation techniques; they are data perturbation anonymization, suppression, generalization, etc. In statistical databases, noise addition techniques are used to protect individuals' privacy, but at the expense of allowing partial disclosure, providing information with less statistical quality, and introducing biases into query responses [17].The idea behind noise addition techniques for PPDM is that some noise is added to the original data to prevent the identification of confidential information relating to a particular individual. Generalization is a simple transformation, where the interval representation for the general domains of both numeric and categorical attributes is used to represent the attribute original value[12].The biggest problem faced by the data publishing community is the process of dividing the attributes of the micro data as quasi identifiers, sensitive attributes and the non-sensitive attributes[3].In this paper feature selection method is used to set the quasi identifiers of the datasets and a hybrid perturbation technique is used to perturb it.

The paper is organized as follows Section 1 of this paper gives the introduction. Section 2 gives the literature survey. Section 3 gives the problem definition. Section 4 gives the methodology used. Section 5 gives the dataset information. Section 6 discusses about information gain feature selection technique. Section 7 gives the Data mining algorithms and metrics used. Section 8 gives the experimental results and discussions. Section 9 gives the conclusion.

2. LITERATURE SURVEY

The term privacy-preserving data mining was introduced by Agrawal [4] and Lindell[5]. These papers considered two fundamental problems of PPDM: privacy-preserving data collection and mining a data set partitioned across several private enterprises. Agrawal and Srikant[4]devised a randomization algorithm that allows a large number of users to contribute their private records for efficient centralized data mining while limiting the disclosure of their values; Lindell and Pinkas[14] invented a cryptographic protocol for decision tree construction over a data set horizontally partitioned between two parties. These methods were subsequently refined and extended by many researchers worldwide. Alexandre Evfimievski et.al [5] discusses about the privacy preservation techniques [2] in-order to preserve privacy. This papers discuss about the basic privacy preservation techniques like suppression, summarization, cryptography and randomization. E. Poovammal et.al provides a detailed study on different privacy preservation techniques like data publishing, k-anonymity, l-diversity and a privacy preserving model [6].This paper discusses about the advantages and disadvantages of k-anonymity, advantage of l-diversity over k-anonymity and finally the privacy preserving model which performs privacy preservation based on the data type. If it is numerical data type, transformation is performed by categorical membership values and if categorical by mapping values. Agrawal and Srikant proposed a noise addition technique which added random noise to attribute values in such a way that the distributions of data values belonging to original and perturbed data set were very different. In this technique it is no longer possible to precisely estimate the original values of individual records [7]. Kargupta et al.

questioned the usefulness of additive noise perturbation techniques in preserving privacy. They proposed a spectral filtering technique which makes use of the theory of random matrices to produce a close estimate of an original data set from the perturbed (released) version of the data set.[13] Islam et al proposed a framework for adding noise to all attributes both numerical and categorical in two steps; in the first step following a data swapping technique we add noise to sensitive class attribute values, which are also known as labels[11]. Peng peng Lin et.al in their work have explored the use of feature selection techniques for privacy preservation purpose. Sparsified singular value decomposition is used for data distraction and filter based feature selection method is used for feature selection. Metrics used here are the data distortion levels. The mining utility of the distorted data is tested with SVM classifier. In this work top ranked attributes are selected using best first search method filter method is used using correlation based feature evaluator. [17] The biggest problem faced by the data publishing community is the process of dividing the attributes of the micro data as quasi identifiers, sensitive attributes and the non-sensitive attributes [3].

3. PROBLEM STATEMENT

Setting the quasi identifiers and sensitive attributes in a database has been always an open problem in privacy preserving data mining. Aggarwal.et.al in his work has discussed that the biggest problem faced by the data publishing community is the process of dividing the attributes of the micro data as Quasi identifiers, Sensitive attributes and the non-sensitive attributes[3].In this paper high ranking features which play an important role in data mining algorithms results are selected using information gain feature selection method and these attributes, which can be linked to identify the underlying individual have been set as quasi identifiers of the dataset. High dimensional datasets like adult and census datasets have been used in this work. A hybrid perturbation method is proposed where both generalization and noise additive perturbation is combined for perturbing the values of the quasi identifiers. The utility of the privacy preserved datasets is evaluated by comparing their accuracy on the two functionalities of data mining namely classification and clustering .The maintenance of rank of attribute before and after perturbation is also evaluated.

4. PROPOSED METHODOLOGY

The methodology used to set the quasi identifiers and perturbation in given in figure 1.The quasi identifier for the datasets taken for study is identified using information gain ranker method and privacy preservation perturbation technique is applied on it.

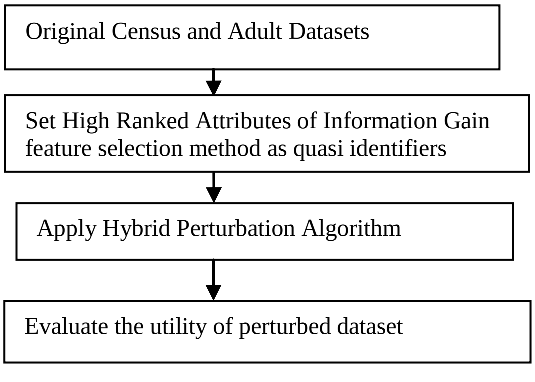

Original Census and Adult Datasets
Set High Ranked Attributes of Information Gain feature selection method as quasi identifiers
Apply Hybrid Perturbation Algorithm
Evaluate the utility of perturbed dataset

Figure: 1Proposed methodology

4.1 Hybrid Perturbation Algorithm

The proposed data perturbation algorithm is as follows:

```
Input: Datasets
Output: perturbed quasi identifiers
Attributes ranking based on information gain ranking method; select high ranked attributes;
If selected attribute= Numeric;
Then divide the domain value of the attribute into ranges;
Level-1 perturbation: N1=N (original value) -lowest value of the range to which it belongs; .
Level-2 perturbation:N2=N1+ Random noise.
Else If selected attribute= categorical
Level-1 perturbation: N1= perturb the original value by generalizing the attribute value and assign a number;
Map the number assigned to N1 to a range.
N2= N1- lowest value of the range to which the number assigned to N1
N3=add random noise to added to numerical value of N2.
End;
```

Figure 2: Proposed Algorithm

The privacy preserved dataset thus formed is tested for its utility on data mining functionalities like classification and clustering.

5. DATA SET INFORMATION

The dataset used in this work are Adult and Census dataset available on UCI Machine Learning Repository [7]. Adult dataset predicts whether the income exceeds $50K/yr. It has a size of 3,755KB. It has 32,561 records where each record contains information about a person. There are 14 attributes including one class attribute, that has two categorical values, ">50K" and "<=50K". The non-class attributes are *age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week,* and *native-country*.

The description of adult dataset is given in Table: 1

Table 1.Adult Dataset Description

Dataset	Adult
Attribute Characteristics:	Categorical, Integer
Number of Instances:	48842
Number of Attributes:	14
Missing Values	Yes
No. of classes	2

Census dataset contains weighted census data extracted from the 1994 and 1995 population surveys conducted by the US Census Bureau. It has a size of 50,800KB.The Census data set has 199524 records where each record contains information about a person. It has 35 attributes of the description of census dataset is given in Table: 2

Table 2 Census Dataset Description.

Dataset	Adult
Attribute Characteristics:	Categorical, Integer
Number of Instances:	199524
Number of Attributes:	42
Missing Values	Yes
No. of classes	2

6. INFORMATION GAIN FEATURE SELECTION METHOD

One of the most important components of a decision tree algorithm is the criterion used to select which attribute will become a test attribute in a given branch of the tree. One of the most well-known measures is the information gain the difference between the original information requirements as given in equation 1. [10]

$$\text{Gain (A)} = \text{Info (D)} - \text{Info (D}_a) \qquad (1)$$

The table 3 list the top ranking attributes of adult dataset

Table 3.Top ranking attributes of Adult dataset

Rank	Attribute name	Rank value
1.	relationship	0.1695
2.	marital-status	0.1625
3.	education	0.1186
4.	Age	0.0834
5.	occupation	0.1514
6.	sex	0.0365
7.	native-country	0.0345
8.	Work class	0.0284
9.	race	0.0166

Among the top ranking attributes *Marital-status, Education, Age, occupation*s are set as quasi identifiers of the Adult dataset, Since they can be linked to identify the entity in the record.

Table 4 lists the top ranking attributes of the census dataset using information gain feature selection method. The top ranking attributes of the table , attributes *Age, Education, Class of worker* are set as quasi identifiers taken for data perturbation.

Table 4.Top ranking attributes of Census dataset.

Rank	Attribute	Rank value
1	Age	0.04898
2	Class of worker	0.03284
3	Major industry recode	0.05635
4	Major occupation recode	0.08261
5	Education	0.11174
6	Country of birth mother	0.02509

7. DATA MINING ALGORITHMS AND METRICS USED:

7.1 Naïve Bayesian (NB) Algorithm

This classifier simply computes the conditional probabilities of the different classes given the values of attributes and then selects the class with the highest conditional probability. If an instance is described with n attributes a_i(i=1…n),then the class that instance is classified to a class v from set of possible classes, according to a maximum a Posteriori (MAP) Naive Bayes classifier is as given in equation (2)

$$v=\arg\max P(v_j)^n \pi_{i-1} p(a_i|v_j) \qquad (2)$$

The conditional probability in the above formula is obtained from the estimates of the probability mass function using the training data. The class probability is not used in these experiments, since no prior phoneme distribution information is available, and thus we are implementing Maximum Likelihood (ML) classification. This Bayes classifier minimizes the probability of classification error under the assumption that the sequence of points is independent. [6]

7.2 K-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms and a non-hierarchical approach that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. A very common measure is the sum of distances or sum of squared Euclidean distances from the mean of each cluster. K-Means training starts with a single cluster with its center as the mean of the data. This cluster is split into two and the means of the new clusters are iteratively trained. These two clusters are again split and the process continues until the specified number of clusters is obtained. [5].

7.3 Noise Addition Technique

In statistical databases, noise addition techniques are used to protect individuals' privacy, but at the expense of allowing partial disclosure, providing information with less statistical quality, and introducing biases into query responses. The idea behind noise addition techniques for PPDM is that some noise is added to the original data to prevent the identification of confidential information relating to a particular individual. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of some patterns that are not supposed to be discovered. We categorize noise addition techniques into three groups: (1) data swapping techniques; (2) data perturbation techniques; and (3) data randomization techniques [16].

7.4 Generalization:

The idea of generalizing an attribute is a simple concept. A value is replaced by a less specific, more general value that is faithful to the original sometimes generalization also includes suppression by imposing on each value generalization hierarchy a new maximal element, atop the old maximal element. [19]

7.5 Metrics:

The metrics used to evaluate the proposed technique are Utility and data distortion

7.5.1 Utility Measure

The classification accuracy and clustering accuracy are calculated using the formula given below [8]

$$\text{Classification Accuracy} = \frac{\text{Number of correctly classified instance}}{\text{Total number of instance in that class.}}$$

$$\text{Clustering Accuracy} = \frac{\text{Number of correctly clustered instance}}{\text{The total number of instance in that class.}}$$

Precision = tp/(tp + fp)

Where tp and fp are the numbers of true positive and false positive predictions for the considered class.

Recall = tp/(tp+fn)

where tp and fn are the numbers of true positive and false negative predictions for the considered class. tp + fn is the total number of test examples of the considered class. fn is the total number of test examples of the considered class.

F-measure can calculate by using the following formula given in Equation 2:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \qquad (2)$$

where P and R is precision and recall. β= Recall parameter.[9]

7.5.2 Data Distortion Measure:

The data distortion measure used in this paper is Maintenance of Rank of Features which is calculated as [20]

$$\text{Maintenance of Rank of Features (CK)} = \frac{\text{Number of attributes that keep their rank after perturbation}}{\text{Total number of features perturbed}}$$

8. EXPERIMENTAL RESULTS AND DISCUSSIONS:

The experiments were conducted using Weka simulation software [15]. In adult dataset, the quasi identifiers taken for data perturbation are *Age*, *Marital status*, *Education* and *Occupation*. Here Age is numerical attribute, whereas *Education*, *Marital status* and *Occupation* are categorical attributes. The perturbed values of the attributes are as given in table 5:

Table 5 Perturbed quasi identifiers of adult Dataset

Age	Education	Marital status	Occupation
27	education 31	marital status 20	occupation 22
27	education 33	marital status 26	occupation 27
45	education 45	marital status 36	occupation 48
25	education 32	marital status 23	occupation 25
34	education 47	marital status 35	occupation 33
43	education 48	marital status 40	occupation 42
32	education 39	marital status 34	occupation 38
38	education 44	marital status 30	occupation 39

The distribution of the quasi identifiers Adult dataset before being perturbed is given in Figure3

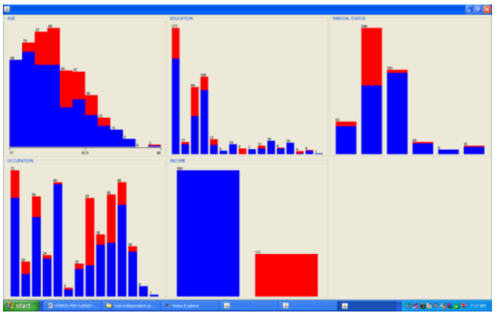


Figure3: Distribution of data in quasi identifiers in adult dataset before perturbation

The distribution of values of the attributes set as quasi identifier of Adult dataset after perturbation is shown in Figure 4

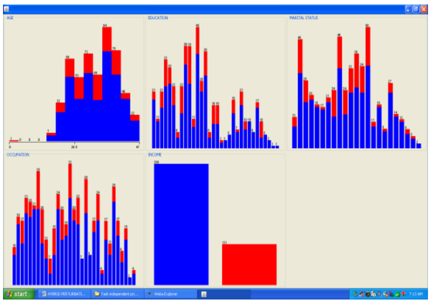


Figure 4: Distribution of data in quasi identifiers after perturbation

By comparing figures 3 and 4 it is incurred that in Adult perturbed dataset, the categorical attributes perturbed have added more labels to them and the numeric attribute the original range 0-90 is distorted to 0-47.

In Census dataset, sensitive attributes taken for data perturbation are *Age*, *Class of worker*, and *Education.* Here *Age*, is numerical attribute, *Class of worker* and *Education* are categorical attribute. The perturbed attribute values are as shown in table 6

Table 6 Perturbed quasi identifiers of Census Dataset

Age	Class of worker	Education
23	class of worker 38	education 46
44	class of worker 8	education 14
48	class of worker 28	education 29
22	class of worker 31	education 42
74	class of worker 42	education 48
31	class of worker 45	education 49
49	class of worker 13	education 21
14	class of worker 29	education 36

The distribution of the quasi identifier before perturbation for the original Census dataset is given in Figure 5.

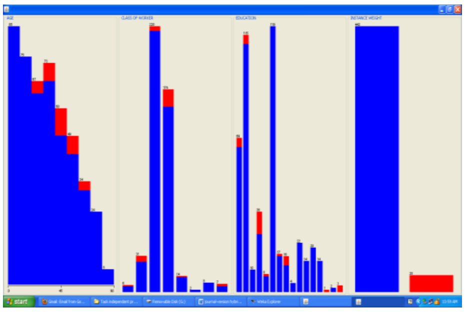


Figure 5: Distribution of data in quasi identifiers of original Census dataset

The distribution of quasi identifier values after perturbation is shown in Figure 6

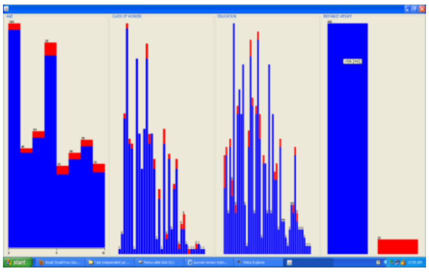
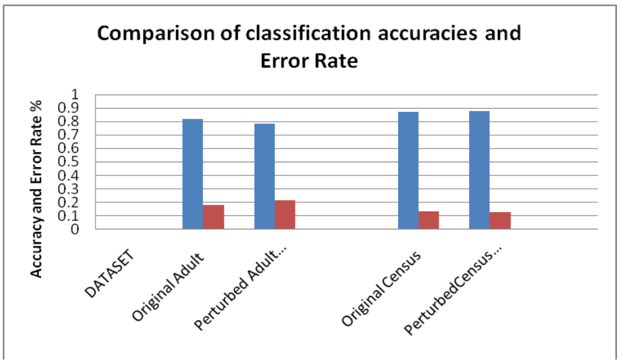


Figure 6: Distribution of data in quasi identifiers in perturbed Census dataset

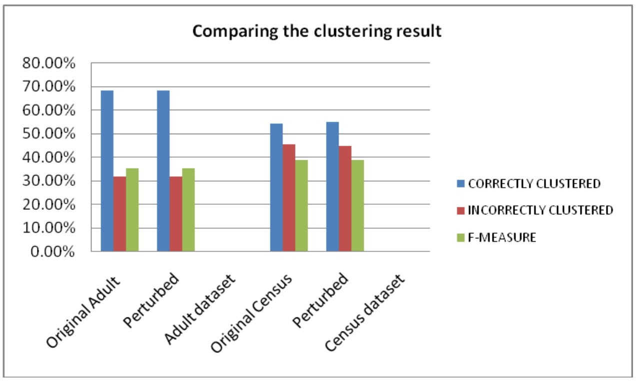
After perturbation the data range in age attribute is modified from 0-90 in the original dataset to 0-10 in perturbed dataset. While the data in the attributes of the class of worker and education have additional labels than the original values.

When the utility factor is considered perturbed Adult and Census datasets are evaluated on its accuracy for Naive Bayes algorithm for classification and the results are as given in Graph 1:



Graph1: Comparison of classification accuracies of original and perturbed datasets

From Graph 1, it is observed that classification accuracy of both original and perturbed Census dataset have the same classification accuracy 81.489%.In Adult original dataset the classification accuracy is82.080%and for the perturbed dataset it is 78.26%, thus change is very nominal. This shows that the utility of the perturbed datasets are comparable with that of the original dataset. The utility of dataset on clustering which is tested using K-Means algorithm and the measures such as correctly clustered instance, incorrectly clustered instance and the F-measure are compared for the original and perturbed for Adult and Census datasets are shown in graph 2.



Graph2: Comparing the clustering result for the Original and perturbed datasets

Graph2 shows that the correctly clustered, the incorrectly clustered and the f-measure values for Census original dataset are 54.4681%, 45.5319% and 0.3898 respectively. For the perturbed dataset the values are 55.1064%, 44.8936%and 0.3898 respectively .Thus, there is very nominal difference in value between the perturbed and original census datasets. When Adult dataset is considered correctly clustered, the incorrectly clustered and the f-measure values for Adult original dataset are 68.2081%,31.7919%and 0.3529respectively for the perturbed dataset the values are 68.2081%,31.7919%and 0.3529respectively.Here both the versions of dataset have given the same result. Thus the hybrid perturbation algorithm protects the privacy as well as the utility of the datasets.

The Maintenance of Rank of Features (CK) of quasi identifiers of two datasets used is shown in table 9

TABLE9. Maintenance of Rank of Features perturbe

Dataset	Quasi-identifiers	Attribute Rank value using info-gain before perturbation	Attribute Rank value using info-gain after perturbation	Maintenance of Rank of Features (CK)
Adult Dataset	Marital status	0.1625	0.0628	0.00
	Occupation	0.1514	0.0588	
	Education	0.1186	0.065	
	Age	0.0834	0.0	
Census Dataset	Education	0.11174	0.07018	0.33
	Class of worker	0.03284	0.07056	
	Age	0.04898	0.0	

The result shows that the CK value for Adult dataset is zero, and hence the quasi attributes of the dataset are fully distorted and hence not able the keep up their rank value. In census dataset out of the three quasi identifiers perturbed, the attribute classes of worker was able to get higher information gain rank value than the original value. The classification accuracy of the perturbed dataset using the proposed hybrid data perturbation technique on both datasets are compared with other privacy preservation techniques like k-anonymization and L-diversity .The comparisons are shown in table 10.

Table10. Comparisons of classification result for adult and census dataset.

Dataset	Privacy preservation techniques applied	Accuracy	Error rate	Correctly classified	Incorrectly classified
Adult	ORIGINAL	82.0809%	17.9191 %	426	93
	K-ANONYMIZATION				
	K=2	80.3468 %	19.6532%	417	102
	K=3	79.1908 %	20.8092 %	411	108
	K=4	78.6127 %	21.3873 %	408	111
	L-DIVERSITY	80.9249 %	19.0751 %	420	99
	DATA PERTURBATION	78.2692 %	21.7308 %	407	113
Census	ORIGINAL	87.0213 %	12.9787 %	409	61
	K-ANONYMIZATION				
	K=2	88.0851 %	11.9149 %	414	56
	K=3	87.4468 %	12.5532 %	411	59
	K=4	88.7234 %	11.2766 %	417	53
	L-DIVERSITY	86.8085 %	13.1915 %	411	59
	DATA PERTURBATION	87.4468 %	12.5532 %	408	62

From the above table it is noted that when the accuracy for classification using NB algorithm is considered in adult dataset, the accuracy decreases by 3-4% as the privacy level increases using k-anonymization .Also for the perturbed dataset the accuracy decreases by 4%. In the case of census dataset the accuracy decreases only by 1-2%. Thus the proposed perturbation method is comparable with other privacy preservation techniques like k-anonymization and L-diversity and the utility of the datasets is also preserved by this proposed privacy preservation technique. The precision and Recall values of the perturbed dataset is compared with k-anonymization and l-diversity techniques on both the datasets on classification using naive Bayes algorithm the values are given in table 11.

Table 11: Precision and recall value for adult and census dataset

Dataset	Privacy preservation techniques Applied	Precision For the class attribute		Recall For the class attribute	
		<= 50K	>50K	<=50K	>50K
Adult	ORIGINAL	0.904	0.623	0.851	0.733
	K-ANONYMIZATION				
	K=2	0.843	0.635	0.902	0.504
	K=3	0.833	0.616	0.902	0.466
	K=4	0.824	0.609	0.907	0.427
	L-DIVERSITY	0.905	0.599	0.832	0.74
	DATA PERTURBATION	0.814	0.46	0.825	0.443
Census	ORIGINAL	0.961	0.211	0.898	0.429
	K-ANONYMIZATION				
	K=2	0.955	0.196	0.916	0.321
	K=3	0.957	0.196	0.907	0.357
	K=4	0.956	0.209	0.923	0.321
	L-DIVERSITY	0.961	0.211	0.898	0.429
	DATA PERTURBATION	0.957	0.196	0.907	0.357

The table 11 shows that the precision values for the class value <= 50K is almost the same on all versions of datasets for both adult and census datasets. But for the class attribute value >50K the precision and recall value decrease and varies about 0.1 to 0.2% from the original datasets value for both the data sets.

The clustering results of various versions of datasets are compared and shown is table 12

Table 12 Comparison of clustering results of adult and census dataset

9	Privacy preservation techniques applied	Correctly clustered	Incorrectly clustered	F-measure
Adult	ORIGINAL	68.2081%	31.7919%	0.3529
	K-ANONYMIZATION			
	K=2	68.2081%	31.7919%	0.3529
	K=3	68.2081%	31.7919%	0.3529
	K=4	68.2081%	31.7919%	0.3530
	L-DIVERSITY	68.2081%	31.7919%	0.3530
	DATA PERTURBATION	68.2081%	31.7919%	0.3529
Census	ORIGINAL	54.4681%	45.5319%	0.3898
	K-ANONYMIZATION			
	K=2	52.766%	47.234%	0.3898
	K=3	51.0638%	48.9362%	0.3898
	K=4	53.8298%	46.1702%	0.3898
	L-DIVERSITY	55.9574%	44.0426%	0.1834
	DATA PERTURBATION	55.1064%	44.8936%	0.3898

The adult datasets results on all the clustering measures like correctly clustered incorrectly clustered and F-measures values for all the versions are same.

In the census dataset F-measure value of l-diversified decrease by 0.2%, but the clustering accuracy is comparable with the original dataset.

9. CONCLUSIONS

The goal of this paper was to set the high ranked attributes selected from information gain feature selection method as the quasi identifier of the dataset, since these attributes are used by data mining algorithms to give patterns and thus knowledge for decision making .The datasets are perturbed using proposed hybrid perturbation algorithm and the fitness was evaluated using Utility and Data distortion measures. The results of the proposed perturbation techniques compared with other privacy preservation techniques like k-anonymity and l-diversity. The results showed that the level of accuracy and hence the utility remained the same for the both original and privacy preserved datasets. The predictive accuracy of the proposed perturbed dataset is comparable with other techniques. When the Maintenance of Rank of Features (CK) measure of quasi identifiers in both the datasets is compared, Census dataset has higher CK value than Adult dataset since the attribute rank value of *class of worker* attribute in Census dataset increased after perturbation. As a future work, efficiency of the perturbation algorithm on various attacks may be tested and other feature selection method like wrapper, gain ratio can be used to set the quasi identifiers of the dataset.

10. REFERENCES

[1] Aggrawal, C.C. (2005*): On* K-Anonymity and the curse of dimensionality. In the proceedings of the 31st conference on VertLargDatabases (VLDB) 901-90.

[2] Agrawal,R.; Srikant,R.(2000): Privacy-Preserving Data Mining by, In Proceedings of the 2000ACM SIGMOD conference on Management of Data, pages 439–450, Dallas, TX, May 14-19 2000 ACM.

[3] Agrawal,R.; Srikant,R.(2000): .Privacy-preserving data mining. In Proc. of the ACM SIGMOD Conference On Management of Data, pages 439-450. ACM Press, May 2000.

[4] Alexandre Evfimievski : Privacy-Preserving Data Mining by IBM Almaden Research Center, USA Tyrone Grandison IBM Almaden Research Center, USA.

[5] Alsabt I.K; Srank ;Singh V. (2006): An Efficient K-Means Clustering Algorithm in 11th International Parallel Processing Symposium, 1998.

[6] Barzan Mozafari,; Carlo Zaniolo.(2006): "Publishing Naive Bayesian Classifiers: Privacy without Accuracy Loss"

[7] Frank, A.; Asuncion, A. (2010): UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[8] Giannella C and Liu K(2009) "On The privacy of Euclidean Distance Preserving Data Perturbation" Computer Science – Cryptography and Security.

[9] Guo, S.Wu, X and Li, Y (2006) "On the Lower Bound of Reconstruction Error for Spectral Filtering based Privacy Preserving Data Mining" in Proceedings of the 10th European conference on Principles and practices of Knowledge discovery in Databases Berlin, Germany.

[10] Han,J.;Kamber,M.(2001): Data Mining Concepts and Techniques, Morgan Kaufmann.

[11] Islam, M.Z.; Brankovic, L.(2007): Privacy Preserving Data Mining: Noise Addition to Categorical Values Using a Novel Clustering Technique, In IEEE Transactions on Industrial Informatics.

[12] Kantarcioglu, M.; Jin, J.; Clifton,C(*2004):* When Do Data Mining Results Violate Privacy? *Proc.* 2004, Int'l Conf. Knowledge Discovery and Data Mining, pp. 599-604.

[13] Kargupta,H.;Datta,S.;Wang,Q.; Sivakumar, K.(2005): .Random-data perturbation techniques and privacypreserving data mining Knowledge and Information Systems, 7:387-414.

[14] Lindell,Y.; Pinkas,B.(2000): Privacy Preserving Data Mining by,In Advances in CryptologyCRYPTO 2000, pages 36–54. Springer-Verlag, Aug. 20-24 2000.

[15] Mark Hall.; Eibe Frank,; Geoffrey Holmes,; Bernhard Pfahringer,; Peter Reutemann,; Ian H. Witten (2009): The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1

[16] Muralidhar,K.; Parsa,R;Sarathy,R.(1999): A general additive data perturbation method for database security.

Management Science, 45(10):1399-1415.

[17] Pengpeng Lin,; Jun Zhang,; Ingrid St. Omer,; Huanjing Wang,; JieWang Proceedings(2011: *A* Comparative study on Data perturbation with feature selection, The international multi conference of Engineers and computer scientist 2011 vol 1, March 16-18, 2011 Hong Kong.

[18] Poovammal,E.; Ponnavaikko,M. (*2009*): Task Independent Privacy Preserving Data Mining on Medical Dataset in 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies.

[19] Sweeney, L. (2002): Achieving k-anonymity privacy protection using generalization and suppression,International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, vol. 10, no. 5, pp. 571,588.

[20] Wang,J,; Zhong,W.J.;Zhang,J.; Xu,S.T.(2006): "Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation," In Proceedings of the 2006 International conference on Information & Knowledge Engineering, pp: 114 - 120, CSREA Press, Las Vegas, Nevada, USA, June 26-29, 2006.