International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

# PRIVACY PRESERVING ANALYSIS OF CENSUS DATASET USING PERTURBATION

[1]***Sangavi N***, *PG Scholar, Bannari Amman Institute of Technology, Sathyamangalam*

[2]***Dr.K.Premalatha***, *Professor, Bannari Amman Institute of Technology, Sathyamangalam*

---

**ABSTRACT:** *Data mining has been facing the serious challenge nowadays due to privacy and security concerns. Enormous amount of data has been generated per second all over the world. The data generated has been given to the third party for the data mining and data analytics. Due to this we are in need of privacy preserving data mining. Data Perturbation is one of the common privacy preserving technique. The sensitive features has to be identified and removed from the dataset (eg. name). Quasi identifiers are the common entity which connect the two or more datasets. For example there may be common attribute like age, gender etc in medical and census dataset that allows to know their details of the person. Quasi identifiers has been applied with the data perturbation technique for privacy preserving. The purpose of doing so is to prevent the privacy of individuals against receiving adverse data. In this system, it will analyses and compare the accuracy of the original and perturbed dataset using classification algorithms such as decision tree, random forest, neural network and SVM.*

***Keywords:*** *Privacy preserving data mining, Data Perturbation, Quasi identifiers, Sensitive features.*

## 1. INTRODUCTION

An number of new data mining algorithms have been suggested with the development of computer data storage capabilities. Further and more knowledge from all the social groups can be accessed. The conventional methods of protection of privacy can not do this well, as when they shield confidential information, intelligence of data is prevented from accessing it. Data mining deals primarily with two aspects of privacy security. First, how to guarantee that the information such as data application process does not reveal such as ID card number, name, address etc. Original quasi identifiers and sensitive information is whether it has been updated or removed from the original database. Quasi identifiers are not themselves sensitive values but they are interconnected with other dataset with common entity which identifies a person. The purpose of doing so is to prevent the privacy of individuals against receiving adverse data. The next is how to make application of data more beneficial. The original dataset has been applied with some statistical measures in order to preserve the privacy. An effective data perturbation has been applied with original dataset and then compared with original dataset. It compare and analyses the performance measure using accuracy over the classification algorithms.

## 2. OBJECTIVE

The main objective of this project is to preserve the privacy of the sensitive data that identifies the person who you are. The sensitive data has been given to the third party for the analytics and the data will be insecure. In order to make the data secure and protect, there is a need of privacy preservation in the data.

## 3. DATASET DESCRIPTION

Census dataset was selected for data analysis. This is taken from UCI machine learning database. The attributes of the census data set are age, job class, occupation, capital gain, capital loss, education, marital status, gender, sex, and marriage, school number, work hours, and country of origin. Age, capital gain, capital loss, hours worked are also important attribute.

## 4. RELATED WORK

### PRIVACY PRESERVING DATAMINING

Data analysis and the techniques of data mining are applicable to many application domains. Any of these areas require handling and frequently publish confidential personal details (e.g. medical records in health care services), which raises concerns about private information being revealed.

Data distribution: Many algorithms do data mining on centralized data, and some on distributed data. Distributed data is vertical and consists of partitioned files. Similar database records in horizontally partitioned data at different sites, and in vertically partitioned data each database records values of attributes at different sites.

Data distortion: Shifting the roots of this method

Data distortion: This technique is to alter the original database record prior to release to achieve privacy security purposes. Data management methods include interruption, blockage, mixing or merging, switching and sampling. Many of these methods are accomplished by altering the value of an attribute or by modifying the value of an attribute by granularity.

| Research Direction | Demonstration |
|---|---|
| General privacy preservation technology | Perturbation Randomization Swapping, Encryption |
| data mining privacy preservation technology | Association Rule Mining Classification, Clustering |
| privacy protection data publishing principle | k-anonymity diversity Invariance Closeness |

**Fig no.4.1 Privacy preserving Techniques**

**Data mining algorithms:** Privacy protection data mining algorithms include classification mining, association rule mining, clustering, and Bayesian networks etc.

**Data or rules hidden:** This approach applies to hiding original data or rules of original data. Since rules hidden from original data are very complex, some people have proposed heuristic method for solving this issue.

Privacy Protection: Data must be carefully updated to achieve a high level of data quality in order to protect the privacy. Do this for as many reasons as.

(1) Change data based on adaptive heuristics methods and only change selected values to mitigate data loss, but not all values.

(2) Encryption techniques, for example efficient multiparty computing. The calculations are secure if each site knows only their input and data but nothing about others.

(3) The method of data reconstruction can reconstruct original distribution of data from random data.

## 5. EXISTING SYSTEM

### DATA PERTURBATION

A procedure for perturbing data can be defined simply as follows. Before the data owners publish their data, they modify the data to mask the confidential information in many ways while maintaining the specific data property that is essential to creating practical data mining models. The intrinsic trade-off between data privacy protection and data utility protection has to be done by perturbation techniques, as disruptive data typically decreases data utility[1].

### Types of Data Perturbation Techniques:

The main additive perturbation technique is the randomisation of additives based on columns.

This type of strategy is based on the reality that

1) Data owners do not want to protect all values in a record equally, so that certain sensitive columns may be skewed by the column-dependent value.

2) The data classification models to be used do not actually require individual records but only a column value distribution meaning separate columns are needed.

CONDENSATION BASED PERTURBATION

- The approach to condensation is a standard multidimensional perturbation technique, aimed at maintaining the matrix of covariance for multiple columns.
- So some geometric properties like decision boundary form are well maintained.
- Unlike the randomization approach, multiple columns as a whole are disturbed in order to generate the whole "perturbed data set."

RANDOM DATA PERTURBATION

- Random projection perturbation refers to the technique of projecting a set of data points to another randomly selected space from the original multidimensional space.
- Let Pk average be a matrix of random projection, where the rows of P are orthonormal[2].

## 6. METHODOLOGY

### Goemetric data perturbation

Geometric data perturbation consists of a sequence of random geometric transformations, including binning, rotation transformation(2 D)

**G(X) = Adding average by Binning +RX (2D)**

i) Binning
- Arrange the attribute in ascending order
- Find the average of the 100 items
- Add the average values to the first 100 items and next average to the next 100 items and so on
- It has been applied to the attribute age and hours of week.

ii) Rotation translation
- A more challenging transformation is rotation
- Rotation of a point by angle in a discrete 2D space is achieved using the transformation matrix

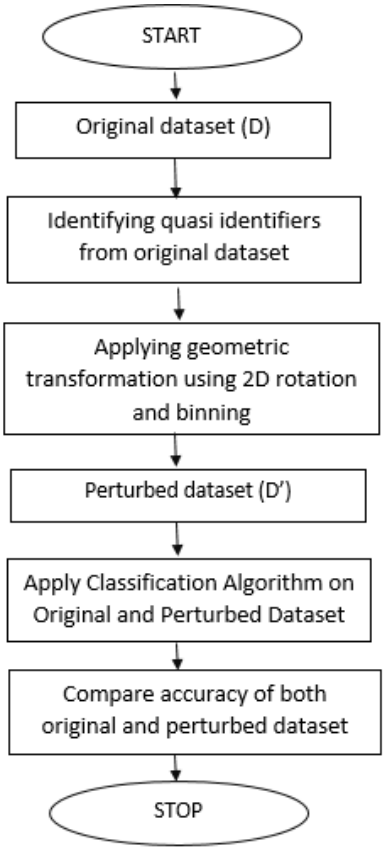


Figure 6.1 work flow diagram

**Procedure:**

Input: Original Data Set D

Output: Comparsion of accuracy with Original Datasets D and Perturbated Dataset D' using classification algorithms.

Step 1: Given input data set D with tuple size n and remove Sensitive Attributes [S]

Step 2: Find the quasi identifiers i.e common entity of two datasets which will identify the person.

Step 4: Apply geometric data perturbation to that attributes

Step 5: Then Perturbated dataset has been done by perturbation technique.

Step 5: Apply decision tree classification algorithm, random forest model, SVM and Linear model on Perturbed dataset D' and original dataset D.

Step 6: Compare and evaluate the results of perturbated D' and original dataset D to assess the accuracy with the classification algorithms.

**CLASSIFICATION ALGORITHMS**

**DECISION TREE**

The leaf node is internal node that is dataset attribute. The root nodes are those groups that either aim 0 or 1. The conditional probability in respect of nodes will be verified. This model provides interpretation of the tree with regard to the attributes [3].

**RANDOM FOREST**

Random forest algorithm is a many of decision trees. In random forests the implementation over the decision tree is mainly used. Implementation of the decision tree is simple when contrasted with random wood. This builds the model and shows the error and time it took to implement the model. It will not require more than 32 rates of the categorical variables. The country attribute has more than 32 categorical values in our dataset [4].

**SVM**

SVM is a linear regression used to characterize the hyper-plane groups. In this model, it separate two sets by using an hyper plane. It calls the tuning factors by kernel, edge gamma, and regularization.. This is a supervised technique of learning which separates the two classes by the hyper plane. It calls certain tuning parameters. They are kernel, gamma, edge, and regularization. The tuning parameters in SVM can help to prevent misclassification. Regularization can quickly maximize smaller margins. Kernel helps help vector machine to solve the equation. Model error is given as relative error. [6].

**NEURAL NETWORK MODEL**

Comparing with other models it is a little complex. The architecture of the neural network is derived from biology i.e. neurons inside our brain. It processes the information in parallel, and how neurons function in the brain. Processing knowledge inside brain is simple for humans. But the knowledge inside the brain is difficult to articulate. One use of the neural network for object detection is the optical character recognition. It allows transferring the information to the brain [7].

**7. EVALUATION OF MODELS**

**ERROR MATRIX CALCULATION**

It should calculate the accuracy of the classification models by error matrix

i) Decision tree:

Decision tree's total error rate is 16.3% and the average class error rate is 27.5% is same in both original and pertubated dataset

```
Error matrix for the Decision Tree model on adult1.csv [test] (counts):

       Predicted
Actual  <=50K >50K Error
  <=50K  3502  217   5.8
  >50K    579  587  49.7

Error matrix for the Decision Tree model on adult1.csv [test] (proportions)

       Predicted
Actual  <=50K >50K Error
  <=50K  71.7  4.4   5.8
  >50K   11.9 12.0  49.7

Overall error: 16.3%, Averaged class error: 27.75%

Rattle timestamp: 2019-09-25 14:47:12 SANGU
```

Figure.7.1 Decision tree original error matrix

```
Error matrix for the Decision Tree model on adult_perturb.csv [test] (counts):

      Predicted
Actual    0    1 Error
     0 587  579  49.7
     1 217 3502   5.8

Error matrix for the Decision Tree model on adult_perturb.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 12.0 11.9  49.7
     1  4.4 71.7   5.8

Overall error: 16.3%, Averaged class error: 27.75%

Rattle timestamp: 2020-03-16 10:12:47 SANGU
```

Figure.7.2 Decision tree Perturbed error matrix

ii) Random forest

The overall rate of random forest error rate of original and perturbated is 14.1% and 14.9% and the average rate of error in the class is 21.5% and 22%

```
Error matrix for the Random Forest model on adult.csv [test] (counts):

      Predicted
Actual    0    1 Error
     0  692  445  39.1
     1  225 3231   6.5

Error matrix for the Random Forest model on adult.csv [test] (proportions)

      Predicted
Actual    0    1 Error
     0 15.1  9.7  39.1
     1  4.9 70.3   6.5

Overall error: 14.6%, Averaged class error: 22.8%
```

Figure. 7.3 Random forest original error matrix

```
Error matrix for the Random Forest model on adult_perturb.csv [test] (counts):

      Predicted
Actual    0    1 Error
     0  691  446  39.2
     1  235 3221   6.8

Error matrix for the Random Forest model on adult_perturb.csv [test] (proportions)

      Predicted
Actual    0    1 Error
     0 15.0  9.7  39.2
     1  5.1 70.1   6.8

Overall error: 14.9%, Averaged class error: 23%

Rattle timestamp: 2020-03-16 10:12:47 SANGU
```

Figure. 4 Random forest pertubated error matrix

iii) SVM

The overall error rate for SVM is 15.7% and the average error rate for the class is 24.9%

```
Error matrix for the SVM model on adult_perturb.csv [test] (counts):

      Predicted
Actual    0    1 Error
     0  646  491  43.2
     1  231 3225   6.7

Error matrix for the SVM model on adult_perturb.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 14.1 10.7  43.2
     1  5.0 70.2   6.7

Overall error: 15.7%, Averaged class error: 24.95%

Rattle timestamp: 2020-03-16 10:12:51 SANGU
```

Figure. 7.6 SVM original error matrix

```
Error matrix for the SVM model on adult1.csv [test] (counts):

       Predicted
Actual  <=50K >50K Error
  <=50K  3160  227   6.7
  >50K    482  634  43.2

Error matrix for the SVM model on adult1.csv [test] (proportions):

       Predicted
Actual  <=50K >50K Error
  <=50K  70.2  5.0   6.7
  >50K   10.7 14.1  43.2

Overall error: 15.7%, Averaged class error: 24.95%
```

Figure. 7.7 SVM pertubated error matrix

iv) Neural network model

The overall error rate of original dataset and perturbated dataset for the neural net is 22.1% and 16%.

Figure.7.8 Neural network original error matrix

```
============================================================
Error matrix for the Neural Net model on adult.csv [test] (counts):

     Predicted
Actual    0    1 Error
     0  192  945  83.1
     1   73 3383   2.1

Error matrix for the Neural Net model on adult.csv [test] (proportions):

     Predicted
Actual    0    1 Error
     0  4.2 20.6  83.1
     1  1.6 73.7   2.1

Overall error: 22.1%, Averaged class error: 42.6%

Error matrix for the Neural Net model on adult_perturb.csv [test] (counts):

     Predicted
Actual    0    1 Error
     0  667  470  41.3
     1  262 3194   7.6

Error matrix for the Neural Net model on adult_perturb.csv [test] (proportions):

     Predicted
Actual    0    1 Error
     0 14.5 10.2  41.3
     1  5.7 69.5   7.6

Overall error: 16%, Averaged class error: 24.45%

Rattle timestamp: 2020-03-16 10:12:52 SANGU
```

Figure. 7.9 Neural network perturbed error matrix

**Accuracy**

Accuracy is the performance measure used to calculate the accuracy rate by using error matrix using the values of true negative, false positives, and false negative values

$$.= \frac{TP+TN}{TP+TN+FP+FN}$$

Where TP is positive, TN is negative and FN is negative

## 8. EXPERIMENTAL RESULTS

The original and perturbed dataset has been applied with classification algorithms such as random forest, decision tree, neural network and svm. Then the accuracy of perturbed dataset is same as original dataset in decision tree and slightly lower in other

Table 1. Accuracy of Perturbed Dataset over classification Algorithms

| | True positive | True negative | False positive | False negative | Accuracy |
|---|---|---|---|---|---|
| Decision tree | 587 | 3502 | 579 | 579 | 83 |
| | | | | | |
| Random forest | 691 | 3231 | 446 | 225 | 82 |
| SVM | 646 | 3230 | 491 | 226 | 84 |
| Neural network | 196 | 3383 | 945 | 945 | 77 |

algorithms. Goemetric Data perturbation works better when comparing with other perturbation

Table 2. Accuracy of original Dataset over classification algorithms

| | true positive | true negative | false positive | false negative | accuracy |
|---|---|---|---|---|---|
| Decision tree | 587 | 3502 | 579 | 217 | 83 |
| Random forest | 691 | 3221 | 446 | 235 | 82 |
| SVM | 646 | 3225 | 491 | 231 | 84 |
| neural network | 667 | 3194 | 470 | 262 | 84 |

techniques. Accuracy is the performance measure used to analyses the original and perturbed dataset by using classification algorithms.

## 9. CONCLUSION

The essential geometric properties can be preserved by geometric disturbance, so most data mining models that search for geometric class boundaries are well preserved with the disturbed data. This provides improved output accuracy and allows use of statistical tests and can easily hold values.

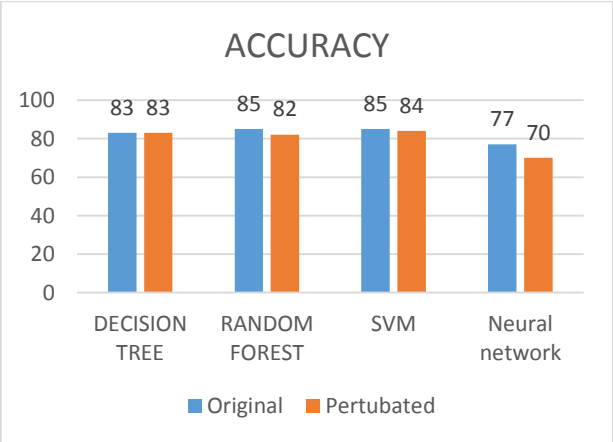


Table .1 Comparison chart

The original and disrupted dataset was implemented with algorithms of classification such as random forest, decision tree, neural network and svm. Then the accuracy of the disturbed data set in decision tree classification algorithms is the same as the original data set and slightly lower in random forest, neural network and SVM classification algorithms

**REFERENCES**

[1]. "Privacy-Preserving Data Mining: Methods, Metrics, and Applications" Ricardo mendes and Joao˜ p. vilela CISUC, Department of Informatics Engineering, University of Coimbra, 3004-504 Coimbra

[2]. "Data Mining: Random Swapping based Data Perturbation Technique for Privacy Preserving in Data Mining" Vasumathi, Ajmeer Khan

[3]. "Census Data Mining and Data Analysis using WEKA", Dr. Sudhir B. Jagtap, Dr. Kodge B. G, International Conference in "*Emerging Trends in Science, Technology and Management*. 2013;10".

[4]. A Comparative analysis of classification algorithms in datamining for accuracy, speed and robustness "Dogan, N. Technol M (2013) 14:105".

[5]. "Survey of Classification Techniques in Data Mining", S. Archana1, Dr. K. Elangovan, *International Journal of Computer Science and Mobile Applications*.

[6]. "Survey on Multiplicative Goemetric Data Perturbation" Aniket Patel1, Keyur Dodiya2, Samir Pate3, International Journal of Research in Advent Technology

[7]. Dr. A. Bharathi, E. Deepan kumar," Survey on Classification Techniques in Data Mining", *International Journal on Recent and Innovation Trends in Computing and Communication*.

[8]. "Credit rating analysis with support vector machines and neural networks: a market comparative study," Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu, *Elsevier Scopus Indexed Journals*.

[9]. "A Comparative Study of Classification Techniques On Adult Data Set" S.Deepajothi, Dr.S.Selvarajan Chettinad college of Engineering and Technology, TamilNadu, India

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

[11]. "The Impact of Data Perturbation Techniques on Data Mining" Rick L Welson, Peter A Rosen January 2002

[12]. "An overview of Multiplicative data perturbation", Keerti Dixit, Bhupendra Pandya, International Journal for Research in Applied Science and Engineeering Technology

[13]. "Geometric data perturbation for privacy preserving outsourced data mining" Keke Chen, Ling Knowledge and Information Systems 29(3):657-695