SUNY POLYTECHNIC INSTITUTE

Data Mining:
Privacy Preservation Using
Perturbation Technique

M.S. Computer Science Project

Nikunjkumar Patel

Project presented to the
Department of Computer Sciences

In Partial Fulfilment of the Requirements for the
M.S. degree in Computer Sciences

Declaration

I hereby declare that the project work entitled "Data Mining: Privacy Preservation Using Perturbation Technique" submitted by me to SUNY POLYTECHNIC INSTITUTE in partial fulfillment of the requirements for the degree of **Master of Science in Computer and Information Sciences** under the guidance of prof. Sengupta, is my own research work and interpretations drawn therein are based on materials collected by myself with proper citations as required and I am solemnly responsible for any errors.

I further declare that, any part of this work has not been submitted and will not be submitted, for the award of any other degree either in this institute or in any other institute, without proper citations.

Date: 05/06/2015

N. R. Patel.

Nikunjkumar R. Patel

U00263435

Data Mining:
Privacy Preservation in Data Mining Using Perturbation Technique

M.S. Project

Department of Computer and Information Sciences

Approved and recommended for acceptance as a project in partial fulfillment of the requirements for the degree of **Master of Science in Computer and Information Sciences**

DATE **05/06/2015**

Prof. Sam Sengupta, Ph.D., Project Advisor

Bruno Andriamanalimanana

Prof. Bruno Andriamanalimanana, Ph. D.

Prof. Jorge Novillo, Ph. D.

Abstract

In recent years, data mining has become important player in determining future business strategies. Data mining helps identifying patterns and trends from large amount of data, which can be used for reducing cost, increasing revenue and many more. With increased use of various data mining technologies and larger storage devices, amount of data collected and stored is significantly increased. This data contains personal information like credit card details, contact and residential information, etc. All these reasons have made it inevitable to concentrate on privacy of the data. In order to alleviate privacy concerns, a number of techniques have recently been proposed to perform the data mining in privacy preserving way.

This project briefs about various data mining models and explains in detail about perturbation techniques. Main objective of this project is to achieve two things. First, preserve the accuracy of the data mining models and second, preserve the privacy of the original data. The discussion about transformation invariant data mining models has shown that multiplicative perturbations can theoretically guarantee zero loss of accuracy for a number of models.

Contents

Chapter 1
Introduction

1.1 Problem Statement

Information has been very important asset for the human life. It gives knowledge and helps understanding things better. In the early years, amount of data was relatively smaller and could be stored using computers and mass digital storage devices. Over the years, amount of data being collected has drastically increased and led to structured databases and database management systems (DBMS). Database management systems are efficient and more importantly can be used effectively for retrieval of particular information from a large collection of data. Data can be in different forms varying from simple numerical measurements and text documents to complex hypertext documents.

Types of data being collected vary from banking transactions, medical and personal data, surveillance video and pictures, digital media, text reports, etc. This information is collected in digital form in databases and/or flat files and used for different purposes. One of the uses is to perform data mining on the given set of data to find out patterns and analyze. Data mining is famously known as Knowledge Data Discovery (KDD). Data mining can be performed on flat files, relational databases, data warehouses, transactional databases, multi-media databases, time-series databases, World Wide Web, etc.

Data mining has both pros and cons. Issues of the data mining include a threat to privacy and security of the data. Sometimes data mining is done by a third party service provider, which compromises the privacy of the data. Motivated by the privacy concerns of the data, new research area called privacy preserving data mining came into picture. Initial idea was to extend traditional data mining technologies by providing mask to sensitive information. But the key issues were how to modify the data and how to get the data stream mining result from the modified data. The goal is to achieve maximum accuracy for the intended data analysis task with less information loss, less response time and maximum privacy gain.

1.2 What is privacy preserving in data mining

In recent years, the growing capacity of information storage devices has led to increased storing personal information about customers and individuals for various purposes. Data mining needs extensive amount of data to do analysis for finding out patterns and other information which could be helpful for business growth, tracking health data, improving services, etc. This information can be misused for many reasons like identity theft, fake credit/debit card transactions, etc. In order to alleviate these concerns a number of techniques have been proposed to perform data mining in privacy preserving way.

Some of the techniques are as following, Data Perturbation (hiding private data while mining patterns), Secure Multi-Party Computation (Building a model over distributed database without knowing others inputs), Knowledge Hiding (Hiding Sensitive rules), Privacy aware knowledge Sharing (do the data mining results themselves violet privacy). In data stream mining, the incoming data is continuous and real time and require algorithm that can change/alter the value before it may become available for data mining using classification or clustering techniques. It`s more difficult to provide privacy to real time data.

1.3 Objectives

The objective of this project is as following.

a) Brief discussion about data mining and data stream mining.
b) Classification of various data mining and data stream mining techniques.
c) Classification of various privacy preserving data mining techniques.
d) Discussion about Multiplicative Data Perturbation.
e) Survey of various Data Perturbation techniques.
f) Study of Geometric data Perturbation Technique and applying it on Data Stream with classification algorithm.
g) Analyze and compare the result of technique with original.

1.4 Goals

The goal is to transform the original data set (S) into modified version of data set (S') that satisfies the given privacy requirement and also preserves as much information as possible for the data analysis task. The data can be continuous or rapid and will be only numeric.

1.5 Organization of Report

Chapter 2 gives brief information about data mining, data stream mining and various techniques used to perform data mining. Chapter 3 discusses about what led to the privacy preservation techniques for data mining. Chapter 4 includes the discussion on the applied method for privacy preservation. Chapter 5 consists of experimental setup. Chapter 6 concludes the project and briefs about future work. Appendix A has code written in Java and the output of the code.

Chapter 2
Literature Survey

2.1Introduction to Database Mining:

“Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).”[2]

Modern computer systems are capable to save at an almost unimaginable rate and from wide range of data generating sources from point of sale machines to machines logging every single credit/debit card transactions, cash withdrawal and check clearances to, to space observing satellites. Here are some examples for better understanding of the volume of the data.

“The current NASA earth observation satellite generates a terabyte of data every data.”[1]

“The human genome project is storing thousands of bytes for each of several billion genetic bases.”[1]

Along with gained storage capacity, scientific research laboratories and business organizations has to come to realize that such massive data holds information which can be critical to business growth or decline, information which could enable accurate predictions of weather and natural disasters, or could lead to cause and possibly to the cure for diseases.

Data mining is also known as Knowledge Discovery in Databases (KDD). While data mining and KDD are often treated as synonyms, data mining is actually a part of KDD process. Knowledge discovery in databases process consists of following steps.

Data cleaning: in this phase noise data and irrelevant data are removed.

Data Integration: in this phase multiple data sources may be combined in a common source.

Data selection: this step decides the relevant data for analysis is decided and retrieved from collection.

Data transformation: selected data is transformed into forms appropriate for mining procedures.

Data mining: it is a crucial step where mining techniques are applied to extract patterns.

Pattern evaluation: valuable patterns are identified based on provided measures.

Knowledge representation: in a finale phase, discovered knowledge is visually presented to the user. It helps user interpret data mining results.

The following figure shows data mining as a step in knowledge discovery process.

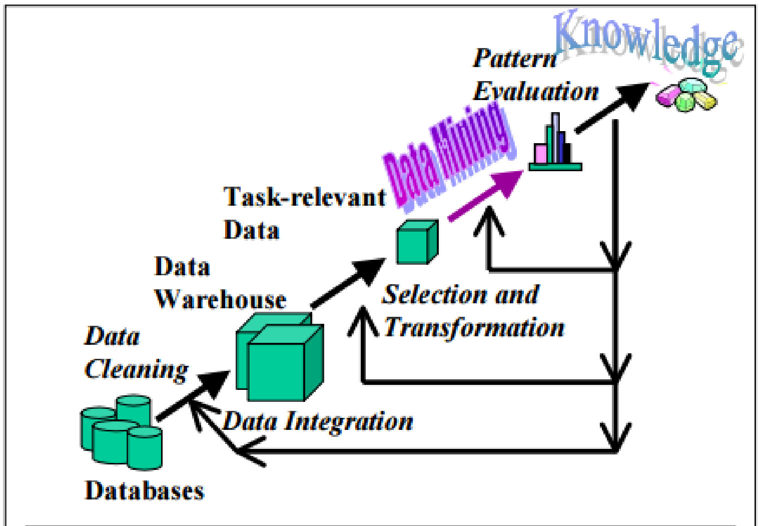


Figure 1.1: Data Mining is the core of Knowledge Discovery process

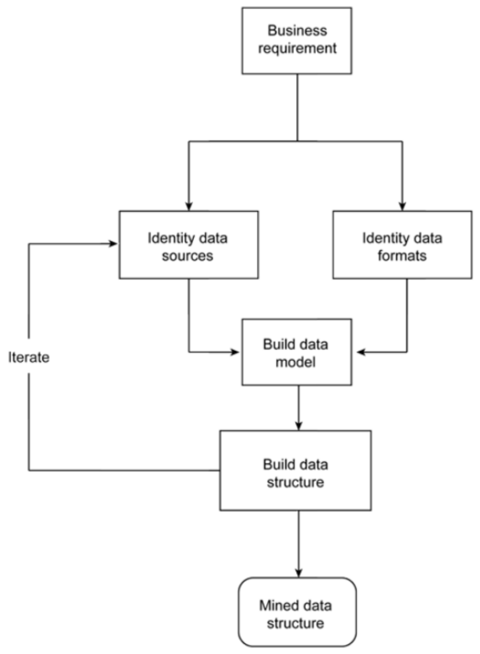
(Figure 2.1 a) [1]

Large data has caused the use of more intense and complex techniques of data mining, partially because the size of data is too big and information varies in its content. With big data sets, merely getting relatively simple and straightforward statistics out of the system is not enough. For example, the country of 100 million populations, knowing that 10 million out of them live in a particular region is not enough. It`s crucial to know their age, economy background, gender, etc. to carry out various government schemes. Similarly, such detailed analysis can come handy for businesses to target potential customers.

Such business-driven needs have changed the simple data retrieval practices into more complex and detailed data retrieval techniques. The process of analysis and knowledge discovery is often iterative as it involves identification of different information that can be extracted.

Data mining also requires the understanding of relating, mapping and clustering the one data to the other data to produce the result.

The figure shown below outlines the process of mining and model building for the business-driven data mining.



(Figure 2.1 b) [3]

2.1.1 Applications of data mining:

- Targeted marketing
- Weather forecasting
- Credit card fraud detection
- Medical diagnosis
- Retail Industry
- Telecommunication industry
- Biological data analysis
- Financial data analysis
- Other scientific application and many more.

Few of many advantages are listed as below.

- Supermarket chains using data mining, collects transactions details to optimize targeting of high value customers and to keep the inventory up to date with the demand.
- A credit card company can detect fraud transactions by keeping track of customer`s transaction patterns and details using data mining
- Predicting viewership for television programs, allowing television executives to schedule/reschedule shows to maximize market share and advertising revenues.

2.2 Data mining Algorithms and Techniques:

Various algorithms and techniques like Artificial Intelligence, Classification, Nearest Neighbor Method, Clustering, Genetic Algorithm, Association Rules, Decision Trees, Regression, Neural Networks, etc., are used for performing data mining based on the desired outcomes, complexity and nature of the data.

a. Association Rule Learning: - This is also called market basket analysis or dependency modelling. It is used to discover relationship and association rules among variables.
b. Clustering: - This technique creates and discovers group of similar data items. This is also called unsupervised classification.
c. Classification: - This can classify data according to their classes i.e. put data in single group that belongs to a common class. This is also called supervised classification. It is most commonly used data mining technique. Classification algorithms work on a basic principle of predicting certain outcome based on a type of input given. “A Classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations.” [7].
d. Regression: - It tries to find a function that model the data with least errors.
e. Summarization: - It provides easy to understand and analysis facility through visualization, reports etc.

Chapter 3
Privacy Preserving in Data mining

3.1Need of Privacy Preservation

Privacy preserving while data mining has been major concern for long time now, with the improvements in technologies, storage devices and advance software, now it is common to have large traditional databases and real time data as well. Few of the main sources of real time data are stock market, weather information coming from satellites, online transactions, internet traffic, telecommunication, etc.

The main difference between traditional databases and real time data is, data in statistical database doesn`t change with time and it can be stored and accessed later. Real time data is a stream of data which needs to be processed within that particular time frame and it is not stored as it is in huge amount. Both category datasets have different approach when it comes to mining.

The main aspect in data mining applications dealing with sensitive information like educational, health care, financial, personal attributes, security, etc. is that it should preserve the privacy of the data. Specially, applications dealing with government records keeping, counter terrorism, self-defense, etc. For example, even though health organizations are allowed to release data after removing the identifiers like name, Identity numbers, address, etc. it is not considered safe enough since re-identification attacks have emerged which can link different public datasets to relocate the original subjects.[10] Sometimes organizations or private entities are not willing to distribute the sensitive information and also patterns detected by data mining systems can be used in a manner which violates the privacy of the individuals or organizations. These privacy constraints have led to exploring the new data mining techniques which protects the privacy of the data and also maintains the efficiency.

3.2Privacy Preserving data mining techniques

There are two dimensions of Privacy preserving in data mining.

Individual Privacy Preservation: This dimension mainly focuses on the privacy of the individuals or private entities.

Collective Privacy Preservation: As the name suggests, main area of this dimension is on the privacy of the overall organizations.

Various techniques have been proposed varying from entire dataset modification to selective dataset modification. Most of privacy preserving data mining techniques can be classified in the following categories

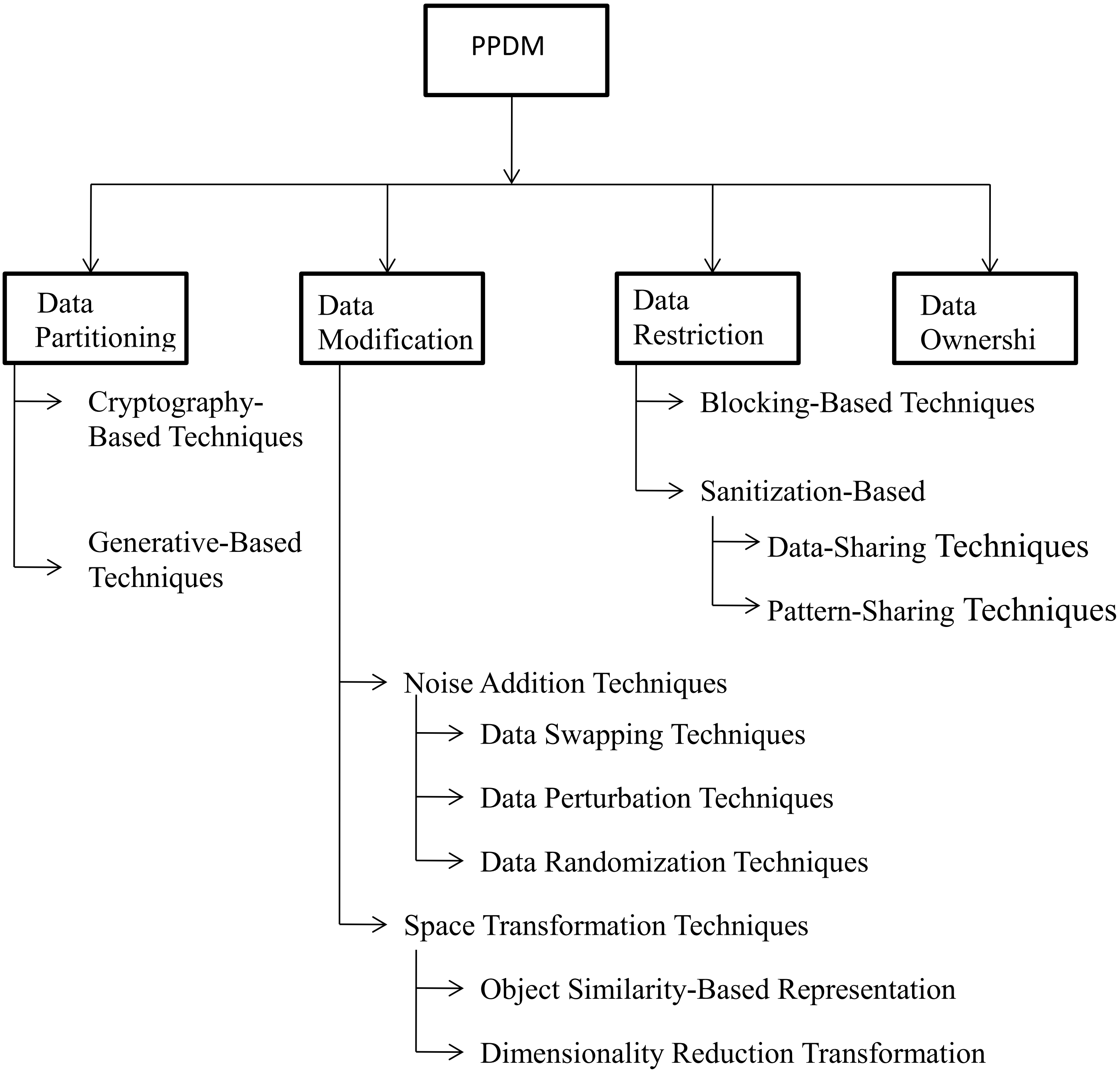
Data Partitioning:

Data Modification:

Data Restriction:

Data Ownership:

Following figure lists out few of privacy preserving data mining techniques practiced by various organizations dealing with the sensitive data.


PPDM
Data Partitioning
Data Modification
Data Restriction
Data Ownershi
Cryptography-Based Techniques
Generative-Based Techniques
Blocking-Based Techniques
Sanitization-Based
Data-Sharing Techniques
Pattern-Sharing Techniques
Noise Addition Techniques
Data Swapping Techniques
Data Perturbation Techniques
Data Randomization Techniques
Space Transformation Techniques
Object Similarity-Based Representation
Dimensionality Reduction Transformation

(Fig. 3.2.1 Privacy preserving data mining techniques)

The focus of this project is on Noise addition techniques, more on data perturbation techniques.

3.3Data Perturbation Techniques

"In recent years, plenty of work has been reported on perturbation techniques for privacy preservation in data mining. Perturbation techniques include random noise addition methods, rotation perturbation, projection perturbation and k-anonymization model. Project primarily focuses on the Geometric data perturbation which can be classified as one of the random noise addition methods."[8].

A perturbation based approach for the privacy preservation relies on two facts.

Users are not equally concerned about the privacy of all attributes in the records. Hence, users may be willing to provide those certain values with some modifications with the help of available data perturbation techniques. These modified values can be generated using coding with the help of various programming languages like Java, .Net, C++, etc. or using browser plug in tool [13].

Data mining problems do not necessarily need individual records but only distributions. Since the perturbation distributions are known, it can be possible to reconstruct aggregate distributions and this aggregate information can be used for the data mining algorithms [13].

Any Perturbation technique is evaluated mostly on two bases: the level of privacy preservation and the level of preserved data utilization. Though both are conflicting goals, main goal of any perturbation technique is to ensure maximum privacy preservation as well as maximum preserved data utilization.

Data Privacy: Date privacy is generally identified as a level of difficulty attacker has to face in estimating the original data from the perturbed data. For a given perturbation technique, the more difficult estimation of original values, and the higher level of privacy that technique provides. Geometric data perturbation provides moderate level of data privacy but is more efficient compared to other algorithms [8].

Data Utility: the level of data utility refers to the amount of critical information preserved after perturbation.

"More specifically, the sensitive information is to be model or task oriented. For example, decision tree and k-nearest Neighbor classifier for classification modeling utilizes different sets of information about the datasets. Decision tree construction primarily concerns the related columns of distributions; the k-nearest neighbor relies on the distance relationship involving all the columns. It is interesting to note that data privacy level enhancement share contradictive relationship with data utilization in most of the data perturbation techniques. Mostly perturbation algorithms aiming at maximizing privacy preservation have to bear with less data utility. This innate correlation between both factors makes it challenging for any data perturbation techniques to find a balance [8].

Some of the data perturbation techniques are mentioned as following.

Noise Additive Perturbation: “It is a column based additive randomization. This kind of techniques is based upon the above mentioned two facts that 1) data owner does not want to protect all values in a record equally, this gives the freedom to apply column based distortion on some sensitive fields. 2) Classification models to be used do not require the individual records but column value distributions assuming that they are independent columns [8]. The basic method adds the certain amount of noise to the columns, keeping the structure intact and can be easily reconstructed from perturbed data.

A typical random noise addition model can be described as following. Let a variable X having some distributions, be described as (x1, x2, x3... xn). The random noise addition process changes its original value by adding some kind of noise R and generates perturbed value Y. Now Y will be X + R, resulting into (x1+r1, x2+r2, x3+r3 …xn+rn). Using this noise R, the original value X can recovered by applying reconstruction algorithm on the perturbed values [8].

While this approach is simple, it has some cons. Several researchers have found that it is easy to perform reconstruction based attacks, which is major weakness of randomized noise addition approach. Also resembling properties of the perturbed data can become handy to identify and remove noise from the perturbed data. Moreover, algorithms like decision tree algorithms and association rule mining algorithms, meeting the assumption of independent columns and work on column distributions, can only be modified to reconstruct the column distributions from perturbed datasets” [8].

Condensation-based algorithm: “This approach is a multi-dimensional data perturbation technique, aiming at preserving the dispersion matrix for multiple columns. Some geometric properties like shape of the decision boundary are preserved well. This algorithm unlike the randomize approach, perturbs multiple columns as a whole and generates the entire dataset. Because of above mentioned properties, many existing data mining algorithms can be directly applied to the perturbed dataset without any change or need to develop new algorithms” [8].

“The approach can be described briefly as follows. Algorithm starts by partitioning the original dataset D into number of record groups, say k-record groups. Each group consists two parts. One is a center of the group, selected randomly from the original dataset and the other part is of (k-1) members from original dataset, found using k-1 nearest neighbors. These selected k records are first removed from the original dataset before forming the next group. Advantage of having small locality of the group, it is possible to regenerate a set of k records to preserve the covariance and distribution.

The size of the locality is reciprocal of the preservation of covariance with regenerated k records. If size of locality in each group is smaller, then it offers better quality of covariance preservation for regenerated k records” [8].

Rotation perturbation: “It was initially proposed for privacy preservation in data clustering. It is one of the major components in Geometric data perturbation. Rotation perturbation is defined as G(X) = R*X where Xd×n is the original dataset and Rd×d is randomly generated rotation matrix. Distance preservation is the unique benefit as well as major weakness of this method. This method is vulnerable to distance-inference attacks” [8].

Random Projection Perturbation: “It refers to the technique which projects a set of data points from the original multi-dimensional space to another randomly chosen space. Let Pk×d be a random projection matrix. Where, P’s rows are orthonormal.

$$G(X) = \sqrt{\tfrac{d}{k}}PX$$

The above formula is applied to perturb the dataset C. The rationale of projection perturbation depends on its approximate distance preservation, which is supported by Johnson – Lindenstrauss Lemma. The lemma indicates that any given dataset in Euclidean space can be embedded into another space in a way that pair-wise distance of any two points is maintained with small error, resulting into model quality preservation” [8].

Geometric data Perturbation: It consists of sequence of random geometric transformations including multiplicative transformation(R), translation transformation (Ψ) and distance perturbation (Δ).

$$G(X) = RX + \Psi + \Delta$$

Multiplicative transformation (R): This component can be rotation matrix or random projection matrix. Rotation matrix exactly preserves distances while random projection matrix only approximately preserves distances. A key feature of rotation matrix is preserving Euclidean distance. Rotation perturbation is a key component of geometric perturbation which provides primary protection to the perturbed data from naïve estimation attacks. Other components of geometric perturbation are used to protect rotation perturbation from more complicated attacks. A random project matrix R k×d is defined as $R = \sqrt{\frac{d}{k}}R_0$. . The Johnson-Lindenstrauss Lemma proves that random projection can approximately preserve Euclidean distances if certain conditions are satisfied. [8]

Translation transformation Ψ: For any two points x and y in the original space, with translation the new distance will be || (x-t) – (y-t) || = | x – y ||. Therefore, translation always preserves the distance. Only translation perturbation does not provide protection to the data, it can be simply canceled if the attacker knows that only translation perturbation is applied. Translation combined with rotation perturbation, can increase the overall resistant to the attacks.

Distance Perturbation: the above two components preserve the distance relationship. However, distance preserving perturbation can be under distance- inference attacks. The goal of distance perturbation is to preserve distances approximately, while effectively increasing the resistance to the distance-inference attacks. Here, distance perturbation can be noise. Solely applying noise without applying other two components will not preserve the privacy since noise intensity is low. The major issue of distance perturbation is a tradeoff between reduction of model accuracy and gain of privacy guarantee. The data owner may opt not to use distance perturbation with the assumption that data is secure and attacker does not know about the original data. hence, distance-inference attacks cannot happen.

The below graph will help summarizing about Random rotation, random projection and geometric data perturbation.

Random Rotation	**Geometric Perturbation**	**Random Projection**
Y = R*X X is the original dataset for all three formulas Y is the perturbed dataset for all three formulas R is the random rotation matrix	Y = RX + T + D R is the secret rotation matrix (preserves Euclidean distances) T is the secret random translation matrix. D is the secret random noise matrix.	Y = A*X A is the random projection matrix.
Distances are preserved. Less secured [12].	Distances are approximately preserved [11].	Distance is not well preserved. Loss of Data [11].
Accuracy depends on the rotation matrix	Good accuracy than any other perturbation techniques.	Worse accuracy than Geometric data perturbation

Chapter 4
Applied Method

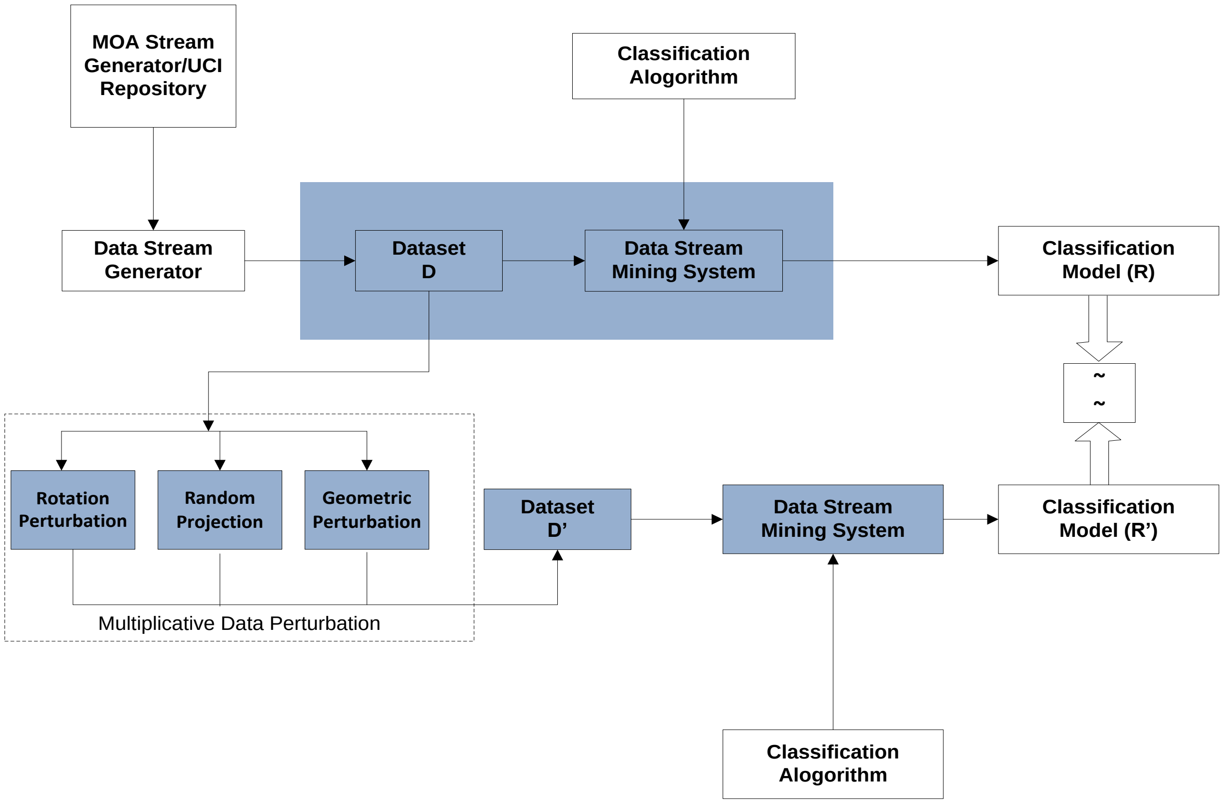
4.1 Introduction:

The method uses Geometric Data Perturbation technique to maintain the privacy of the sensitive data and achieve the maximum efficiency possible. Above mentioned process can be divided into two stages. 1) Dataset preprocessing and 2) data stream classification data mining.

Data preprocessing stage consists of several steps which are discussed in detail in the later chapter. The main objective of first stage is to preprocess the acquired data and generate the perturbed data stream to preserve privacy. The main objective for second stage is to mine the perturbed data streams to classify them. Second stage is simple as it uses the already established algorithms from WEKA tool. The key characteristics of the method applied is, it is simple and easy to implement, lower complexity, requires no complex mathematical calculations. One important thing is time complexity is directly proportional to number of instances to be processed. It means, with growing number of instances or large datasets the time taken also increases.

This technique can be applied to real time data as well as traditional data. Here, traditional datasets are used for experimental purpose.

4.2 Applied Framework


MOA Stream Generator/UCI Repository
Classification Alogorithm
Data Stream Generator
Dataset D
Data Stream Mining System
Classification Model (R)
~
~
Rotation Perturbation
Random Projection
Geometric Perturbation
Multiplicative Data Perturbation
Dataset D'
Data Stream Mining System
Classification Model (R')
Classification Alogorithm

(fig. 5.2.1 Applied Framework for privacy preservation in dataset using classification)

The above graphical representation of the applied framework illustrates privacy preserving process using Geometric data perturbation technique. There are some highlighted areas in the figure, which represents the two different stages of the process.

The top left box of the figure is MOA (Massive Online Analysis) Generator or UCI repository, which are the source of real time data/ data stream or traditional data. Data is then sent to Data stream generator which collects the required information and stores in the database. Dataset D is sent to data mining systems like MOA, WEKA, RapidMiner, R-Programming, Orange, etc. and classification algorithm like is applied. Then dataset D is altered by applying any of Multiplicative Data Perturbation techniques and modified dataset D' is obtained. Now, this classification algorithm is also applied on modified dataset D'. Final step is to compare results of the both datasets.

For now, part of the applied framework will be implemented and remaining can be implemented as a future research work. Current implementation involves traditional datasets instead of real time data. Because of traditional dataset being used, for the algorithm implementation purpose, WEKA tool will be replacing the MOA which is used for real time data or data streaming. Framework will be using Geometric data perturbation as a multiplicative data technique.

4.3 Algorithm: Geometric Data Perturbation

The idea behind using Geometric Data Perturbation algorithm is, because of its simplicity. Geometric perturbation is nothing but the enhancement to the rotation perturbation by coupling it with additional components like random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R £ X. It will be clear that by adding those additional components, Multiplicative perturbation for privacy preserving data mining geometric perturbation exhibits more robustness and provide efficiency in countering the attacks compared to normal rotation based perturbation.

For each attribute of G(X), let T be the translation, random rotation R, D be a Gaussian Noise and X be the original dataset. The value of the attribute G(X) can be found using following formula.

G(X) = R*X + T + D

Procedure: Geometric transformation based Multiplicative data perturbation

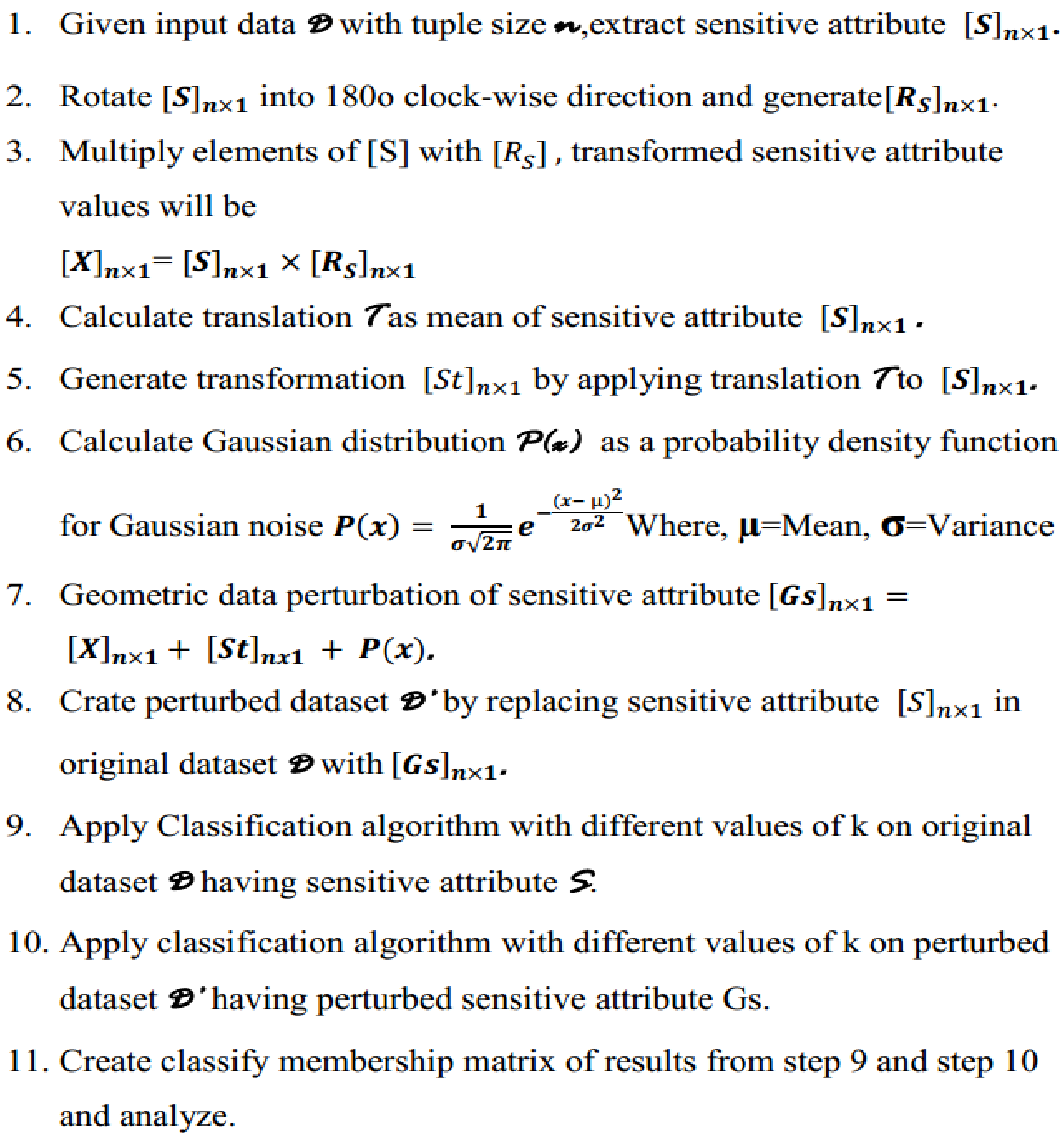
Input: Dataset D, Sensitive attribute S.

Intermediate result: Perturbed dataset D'

Output: Classification result R and R' for dataset D and D' respectively.

Remarks: As of now, will be using traditional dataset instead of data stream and sensitive attributes will be numerical only.

Steps:



1. Given input data $\mathcal{D}$ with tuple size n,extract sensitive attribute $[S]_{n\times1}$.
2. Rotate $[S]_{n\times1}$ into 180o clock-wise direction and generate$[R_S]_{n\times1}$.
3. Multiply elements of [S] with $[R_S]$, transformed sensitive attribute values will be
   $[X]_{n\times1}= [S]_{n\times1} \times [R_S]_{n\times1}$
4. Calculate translation $\mathcal{T}$as mean of sensitive attribute $[S]_{n\times1}$.
5. Generate transformation $[St]_{n\times1}$ by applying translation $\mathcal{T}$to $[S]_{n\times1}$.
6. Calculate Gaussian distribution $\mathcal{P}(x)$ as a probability density function for Gaussian noise $P(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Where, μ=Mean, σ=Variance
7. Geometric data perturbation of sensitive attribute $[Gs]_{n\times1} = [X]_{n\times1} + [St]_{nx1} + P(x)$.
8. Crate perturbed dataset $\mathcal{D}'$by replacing sensitive attribute $[S]_{n\times1}$ in original dataset $\mathcal{D}$ with $[Gs]_{n\times1}$.
9. Apply Classification algorithm with different values of k on original dataset $\mathcal{D}$ having sensitive attribute $\mathcal{S}$.
10. Apply classification algorithm with different values of k on perturbed dataset $\mathcal{D}'$having perturbed sensitive attribute Gs.
11. Create classify membership matrix of results from step 9 and step 10 and analyze.

(fig. 5.3.1 algorithm steps) [9]

As mentioned earlier, Geometric data perturbation G(X) = R*X + T + D.

Computation:

Generation of T (Translation Matrix):

Here, Translation matrix T = (Original value + Mean value)

Where Mean value = (Sum of original data / no. of elements)

For example,

Original dataset(x) = 10, 20, 20, 14, 23.2, 21.4, 40, 46.5

Mean value (M) = (10+20+20+14+23.2+21.4+40+46.5)/8

= (195.1/8) = 24.3875

Translation Matrix (T) = (10+M, 20+M, 20+M, 14+M, 23.2+M, 21.4+M, 40+M, 46.5+M)

= (34.3875, 44.3875, 44.3875, 38.3875, 47.5875, 45.7875, 64.3875, 70.8875)

At the end, all these values of translation matrix will be added to rotated original data.

Rotation of Original data (Rotation Matrix):

Here, rotation matrix R = Rotation of original data to 180 degree

For example,

Original dataset(x) = 10, 20, 20, 14, 23.2, 21.4, 40, 46.5

Rotation matrix (R) = 46.5, 40, 21.4, 23.2, 14, 20, 20, 10

Generation of Noise (D):

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} ,$$

The above equation is for the random Noise D. It is Gaussian Noise in this case.

Equation = [(1/ (variance * sqrt(2*PI))) * (pow (e, (-pow (x-mean), 2)/2*pow (varience, 2)))]

Explanation of variance:

Variance = sqrt (pow ((original dataset (X) - Mean), 2)/no of elements)

For example,

Step 1: pow (take data element first - mean, 2) = pow (1-3, 2) =4

Step 2: pow (take data element second - mean, 2) = pow ((2-3), 2) =1

Step 3: pow (take data element third - mean, 2) = pow ((3-3), 2) =0

Step 4: pow (take data element forth - mean, 2) = pow ((4-3), 2) =1

Step 5: pow (take data element fifth - mean, 2) = pow ((5-3), 2) =4

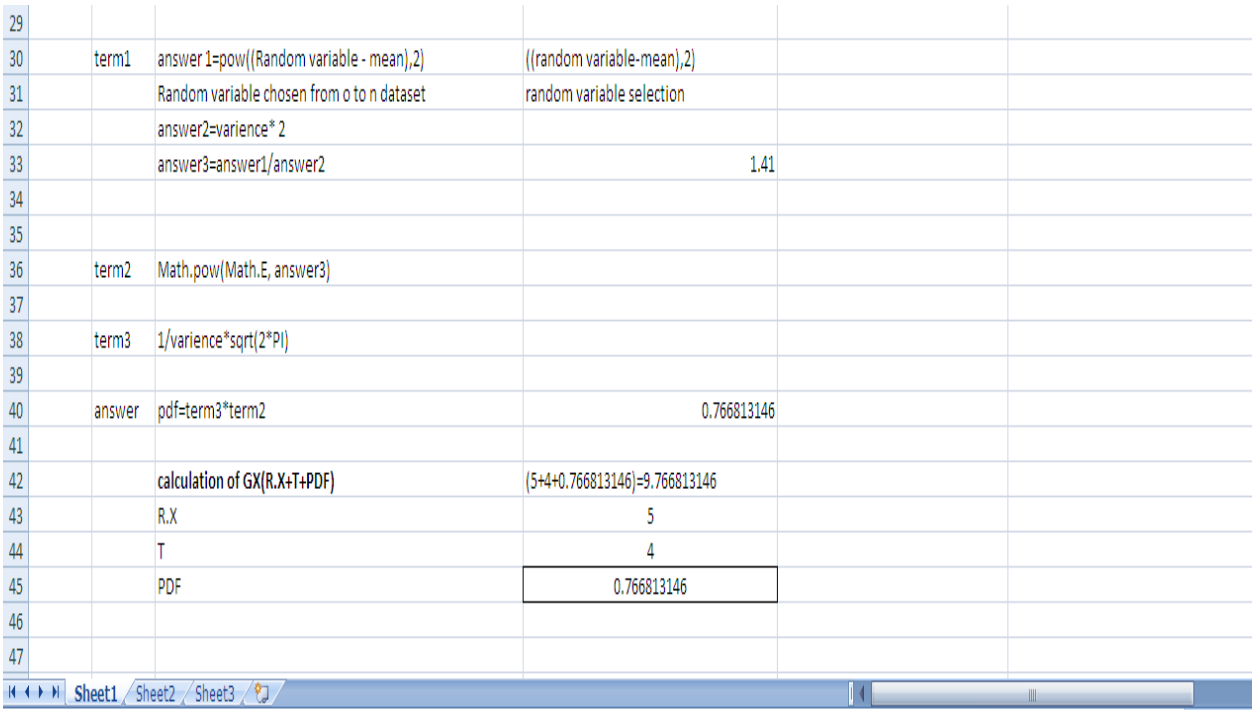
Sum (4 + 1 + 0 + 1 + 4) = 10

Ans = sum/no. of element = 10/5 = 2

Variance = Sqrt (Ans) = sqrt (2) = 1.4142

The below snapshot will show the calculation of Gaussian Noise in the excel sheet.

The following calculation is for the original dataset having values (1, 2, 3, 4, 5)



29				
30		term1	answer 1=pow((Random variable - mean),2)	((random variable-mean),2)
31			Random variable chosen from o to n dataset	random variable selection
32			answer2=varience* 2	
33			answer3=answer1/answer2	1.41
34				
35				
36		term2	Math.pow(Math.E, answer3)	
37				
38		term3	1/varience*sqrt(2*PI)	
39				
40		answer	pdf=term3*term2	0.766813146
41				
42			calculation of GX(R.X+T+PDF)	(5+4+0.766813146)=9.766813146
43			R.X	5
44			T	4
45			PDF	0.766813146
46				
47				

Final G(X) will be as following.

ROTATED	Multiplication of ROTATION and ORIGINAL DATASET	Mean: SUM OF (original)/no of elements	Translation Matrix:(original + mean)	PDF(gaussien noise)	GX(geometirc data perturbation)(R.X+T+PDF)
5	5	3	4	0.766813146	9.766813146
4	8		5		13.76681315
3	9		6		15.76681315
2	8		7		15.76681315
1	5		8		13.76681315

Chapter 5
Implementation

5.1Datasets:

The applied method implements algorithm operations on three different datasets Bank management dataset, Adult dataset and cover type dataset. These datasets can be obtained from UCI machine learning repository. Each dataset has number of attributes, numerical and non-numerical. As of now, algorithm will be applied on numerical attributes only as mentioned previously. After applying Geometric data perturbation technique, it is passed from classification algorithm and then results of both original and modified datasets are compared.

Datasets	**Description**
Bank Management	Total Instances: 45, 210 Attributes: 9
Adult Dataset	Total Instances: 48,842 Attributes: 6
Cover type Dataset	Total Instances: 81,012 Attributes: 8

(Fig. 6.1.1 list of datasets)

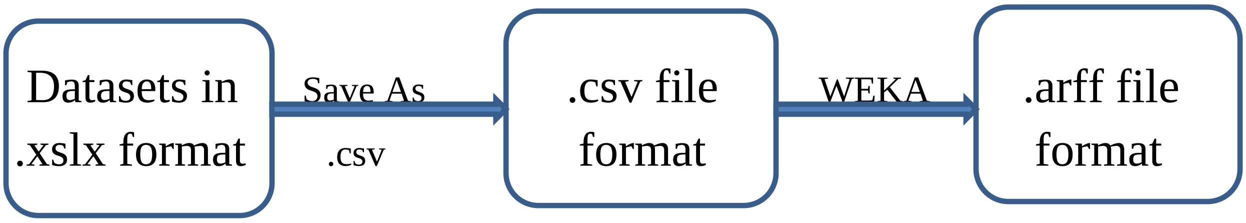
Datasets are required to evaluate data mining process. Datasets like Bank marketing, Cover type, Adult are used for this project.

Attribute	Data type
Age	Numeric
Fnlwgt	Numeric
Work class	Text
Education	Text
Education num	Numeric
Marital Status	Text
Occupation	Text
Relationship	Text
Race	Text
Sex	Text
Capital gain	Numeric
Hours per week	Numeric
Native country	Text
Less greater	Numeric

(Fig. 6.1.2 Adult dataset)

5.2 Preprocessing of Data:

We are going to use MOA (Massive Online Analysis) tool which requires dataset to be in .arff format (Attributes relationship file format). But, datasets obtained from UCI Machine Learning Repository are as Excel files. For converting .xlsx file format into .arff format, first .xlsx file has to be converted into .csv file format and then .csv file has to be converted into .arff format.



File from .xslx format can easily be converted into .csv format using Save As with file type .csv (comma delimited). And .csv can be converted using WEKA tool.

5.3 Steps for Implementation:

Following steps are implemented using Java code. The environment used is Netbeans IDE 8.0.2.

Read the original data from file. (it will be .txt file)

Now rotate the original data to 180°

Perform multiplication of original and rotated data.

Find the Mean value and perform translation.

Find the Gaussian Noise value and print the final result (Final result will be printed into .txt file).

We get perturbed data D'.

Apply WEKA classification algorithms like J248/Naïve Bayesian for original dataset D and perturbed dataset D'.

Compare the results from datasets D and D'.

The example and code with screenshots are attached in the Appendix A for the above mentioned steps.

5.4Experimental Setup:

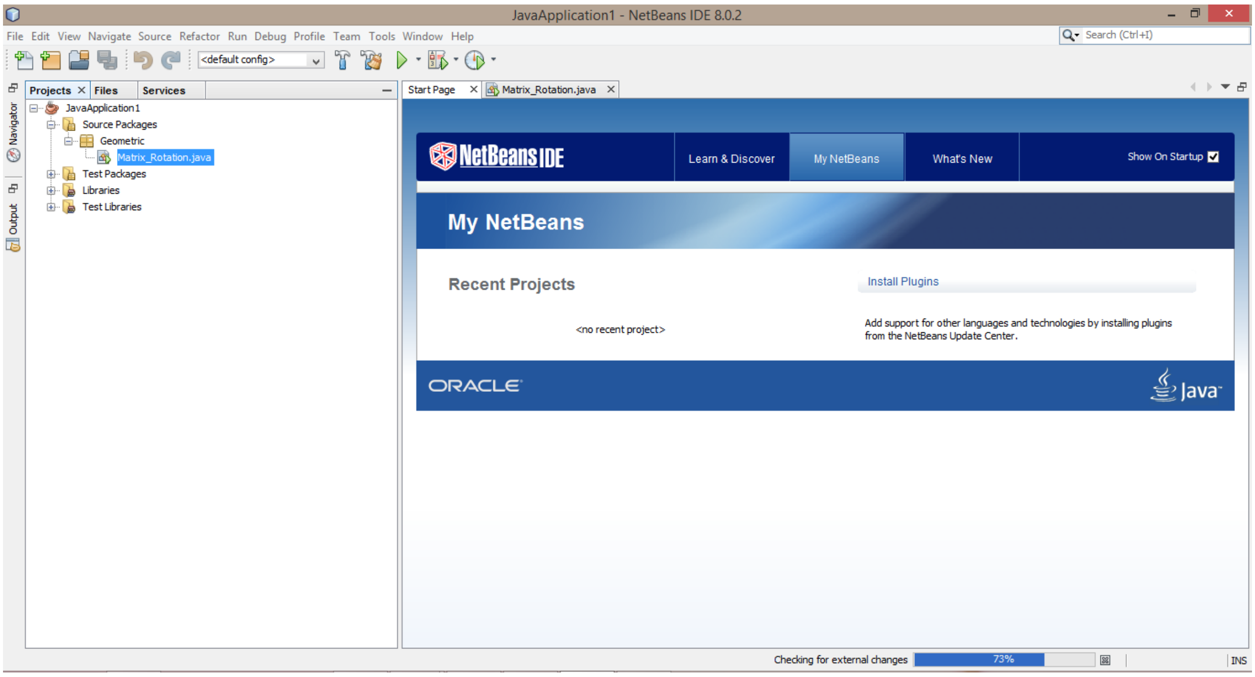
This set up works with attributes having numerical value only and at a time on column only.

For implementation of the Geometric Data perturbation, following environments are used.

NetBeans IDE 8.0.2

"Netbeans IDE is the official IDE for Java 8. It is free and open source and has a large community of users and developers around the world."[7]. It can be downloaded from the following link https://netbeans.org/downloads/.

NetBeans IDE 8.0.2 is a tool used for writing a Java program which adds the Gaussian noise and converts the original dataset D into the perturbed dataset D'. It is difficult to identify original dataset D from perturbed dataset D', unless it is known prior.



<no recent project>

Add support for other languages and technologies by installing plugins from the NetBeans Update Center.

Checking for external changes 73%

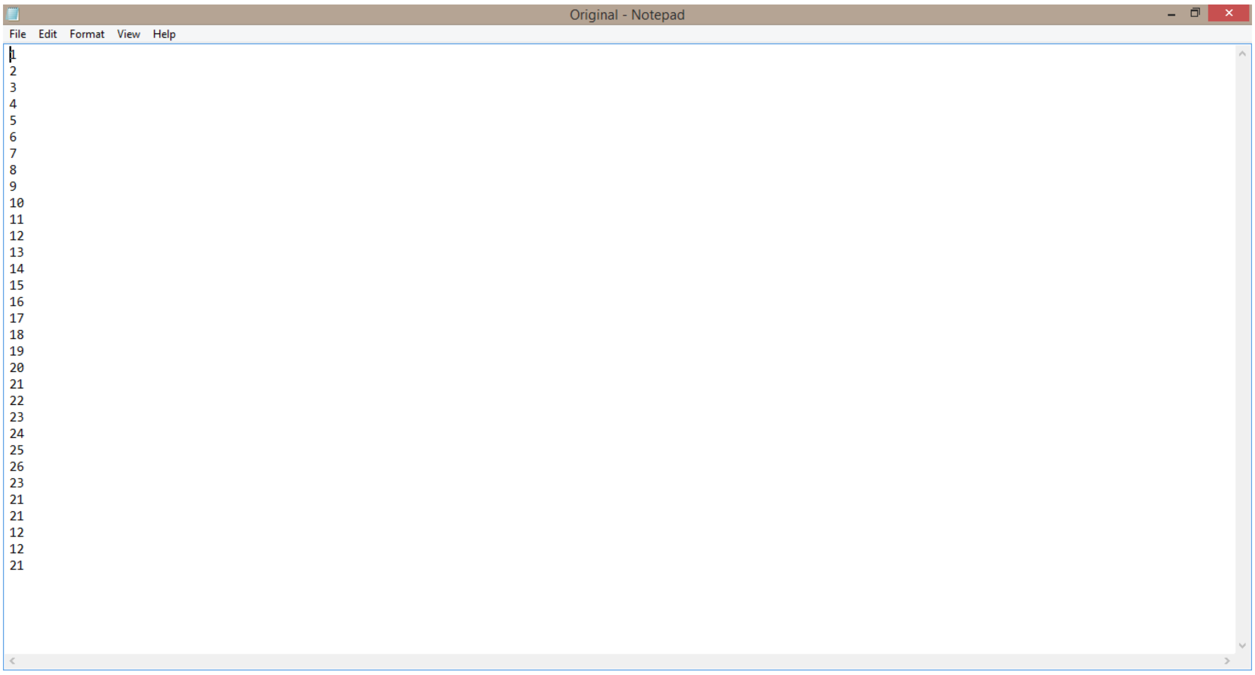
This program is named as JavaApplication1. Netbeans is an open source environment and can be found online.

Notepad

Notepad is a simple text editor which comes with a Windows operating system.

Notepad is used to store numerical attributes of the dataset as a text file. The program which, adds the Gaussian Noise to the original dataset reads dataset from the text file and stores the perturbed dataset into the text file.

Input filename is “Original.txt” and Output filename is “Altered.txt”. Notepad is used to view/modify the dataset.



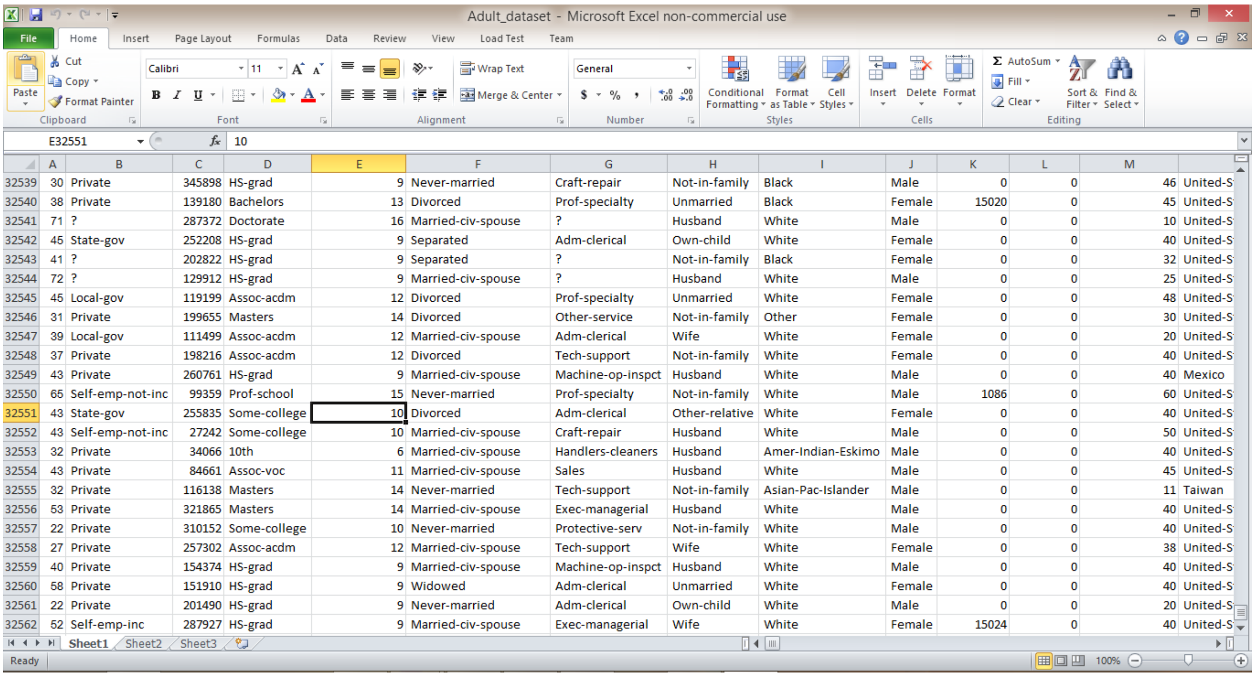
```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
23
21
21
12
12
21
```

Microsoft Excel

The dataset is in .xlsx format. Microsoft Excel is used to view the dataset and then convert it to .csv (comma separated values). Microsoft Excel comes with a Microsoft software package. It is a licensed version, which can be obtained for a cost.

The below mentioned steps will help understanding the conversion from .xlsx format to .csv format.

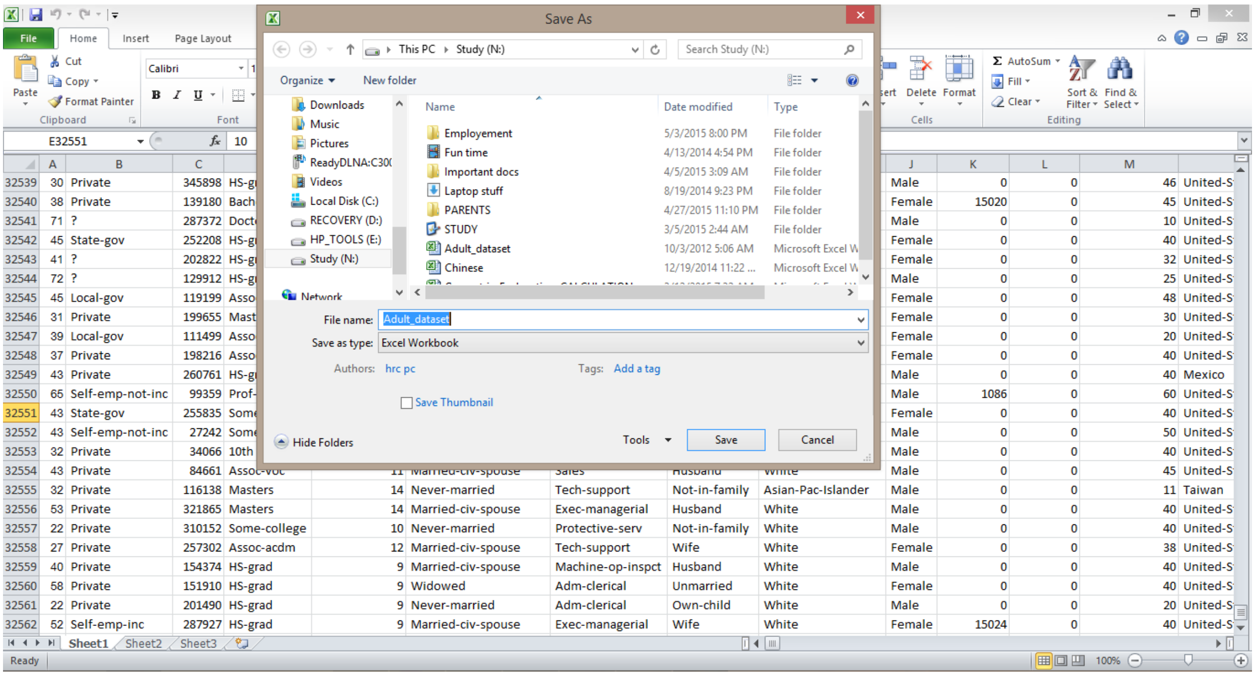
Open the dataset to be converted from .xslx to .csv using Microsoft Excel.



E32551 = 10

	A	B	C	D	E	F	G	H	I	J	K	L	M	
32539	30	Private	345898	HS-grad	9	Never-married	Craft-repair	Not-in-family	Black	Male	0	0	46	United-S
32540	38	Private	139180	Bachelors	13	Divorced	Prof-specialty	Unmarried	Black	Female	15020	0	45	United-S
32541	71	?	287372	Doctorate	16	Married-civ-spouse	?	Husband	White	Male	0	0	10	United-S
32542	45	State-gov	252208	HS-grad	9	Separated	Adm-clerical	Own-child	White	Female	0	0	40	United-S
32543	41	?	202822	HS-grad	9	Separated	?	Not-in-family	Black	Female	0	0	32	United-S
32544	72	?	129912	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	0	0	25	United-S
32545	45	Local-gov	119199	Assoc-acdm	12	Divorced	Prof-specialty	Unmarried	White	Female	0	0	48	United-S
32546	31	Private	199655	Masters	14	Divorced	Other-service	Not-in-family	Other	Female	0	0	30	United-S
32547	39	Local-gov	111499	Assoc-acdm	12	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	20	United-S
32548	37	Private	198216	Assoc-acdm	12	Divorced	Tech-support	Not-in-family	White	Female	0	0	40	United-S
32549	43	Private	260761	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	Mexico
32550	65	Self-emp-not-inc	99359	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	1086	0	60	United-S
32551	43	State-gov	255835	Some-college	10	Divorced	Adm-clerical	Other-relative	White	Female	0	0	40	United-S
32552	43	Self-emp-not-inc	27242	Some-college	10	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	50	United-S
32553	32	Private	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0	0	40	United-S
32554	43	Private	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-S
32555	32	Private	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0	0	11	Taiwan
32556	53	Private	321865	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-S
32557	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0	40	United-S
32558	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-S
32559	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-S
32560	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-S
32561	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-S
32562	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-S

- Click on File -> Save As



Name	Date modified	Type
Employement	5/3/2015 8:00 PM	File folder
Fun time	4/13/2014 4:54 PM	File folder
Important docs	4/5/2015 3:09 AM	File folder
Laptop stuff	8/19/2014 9:23 PM	File folder
PARENTS	4/27/2015 11:10 PM	File folder
STUDY	3/5/2015 2:44 AM	File folder
Adult_dataset	10/3/2012 5:06 AM	Microsoft Excel W
Chinese	12/19/2014 11:22 ...	Microsoft Excel W

File name: Adult_dataset

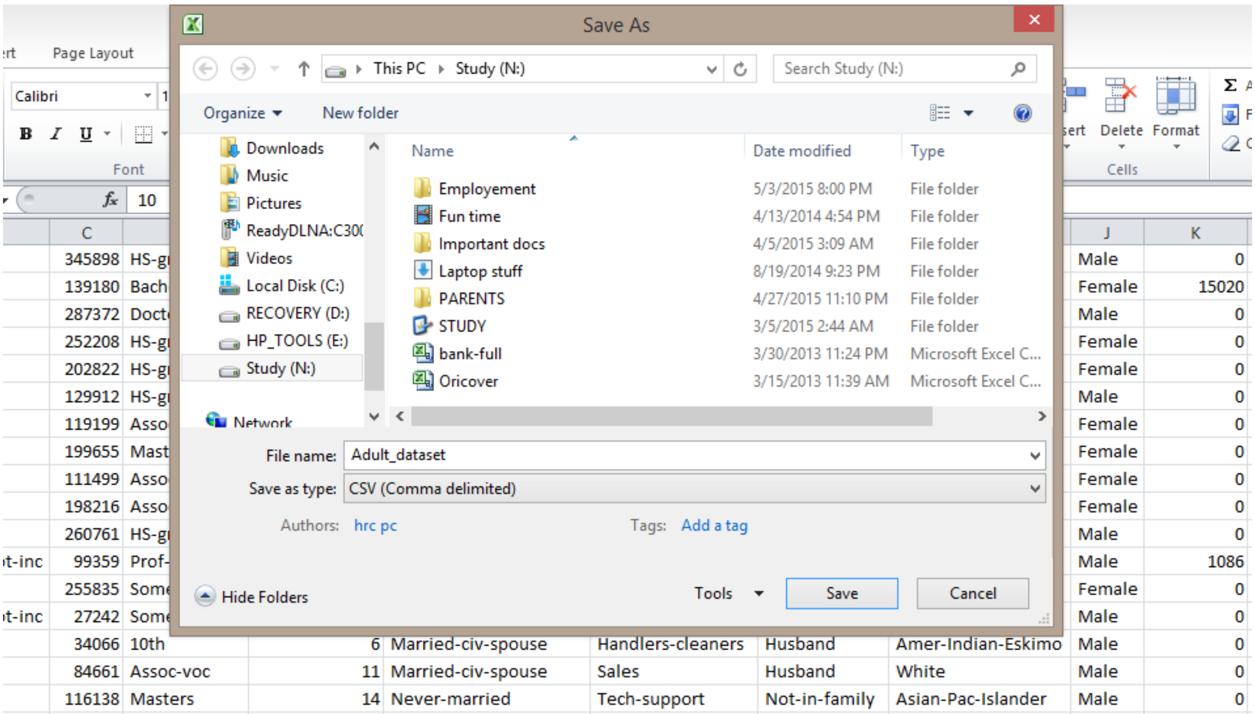
Save as type: Excel Workbook

Authors: hrc pc

Tags: Add a tag

□ Save Thumbnail

- Select Save As Type: CSV (comma delimited) and Click Save.



Name	Date modified	Type
Employement	5/3/2015 8:00 PM	File folder
Fun time	4/13/2014 4:54 PM	File folder
Important docs	4/5/2015 3:09 AM	File folder
Laptop stuff	8/19/2014 9:23 PM	File folder
PARENTS	4/27/2015 11:10 PM	File folder
STUDY	3/5/2015 2:44 AM	File folder
bank-full	3/30/2013 11:24 PM	Microsoft Excel C...
Oricover	3/15/2013 11:39 AM	Microsoft Excel C...

File name: Adult_dataset

Save as type: CSV (Comma delimited)

Authors: hrc pc   Tags: Add a tag

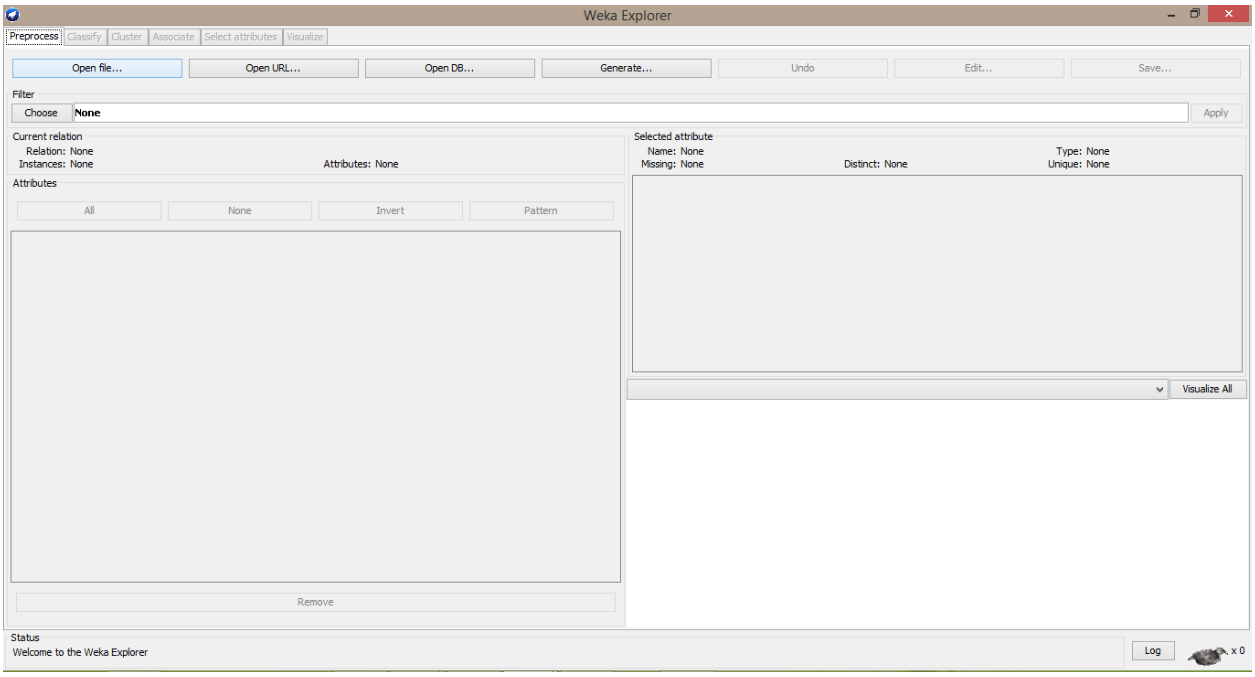
	C							J	K
	345898	HS-gı						Male	0
	139180	Bach						Female	15020
	287372	Docti						Male	0
	252208	HS-gı						Female	0
	202822	HS-gı						Female	0
	129912	HS-gı						Male	0
	119199	Asso						Female	0
	199655	Mast						Female	0
	111499	Asso						Female	0
	198216	Asso						Female	0
	260761	HS-gı						Male	0
ɪt-inc	99359	Prof-						Male	1086
	255835	Some						Female	0
ɪt-inc	27242	Some						Male	0
	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0
	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0
	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0

Weka

“Weka is 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. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Weka is open source software issued under the GNU General Public License.” [5].

Weka is used to convert dataset .csv format to .arff format and then different algorithms are applied on the dataset.

Weka is downloaded from http://www.cs.waikato.ac.nz/ml/weka/downloading.html. This project is using stable 3rd version for Windows X64.



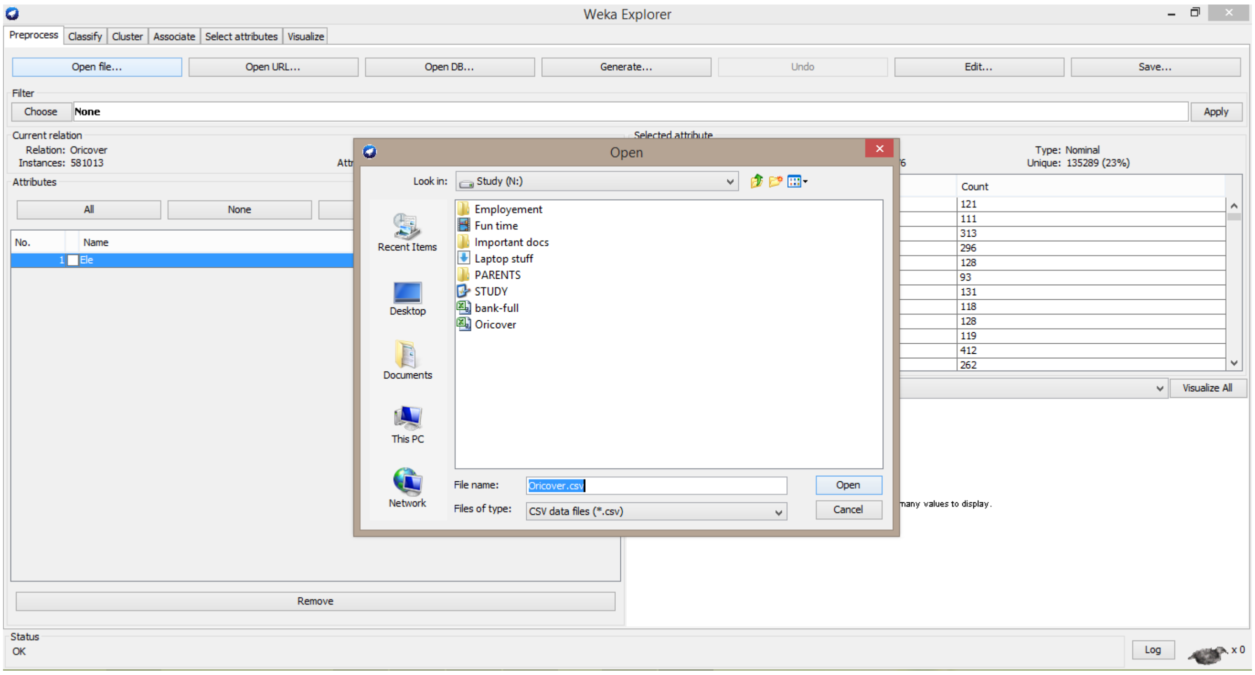
Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

- Click on Open file. Select type as .csv



Filter: None

Relation: Oricover
Instances: 581013

1 Ele

Type: Nominal
Unique: 135289 (23%)

Count
121
111
313
296
128
93
131
118
128
119
412
262

Look in: Study (N:)

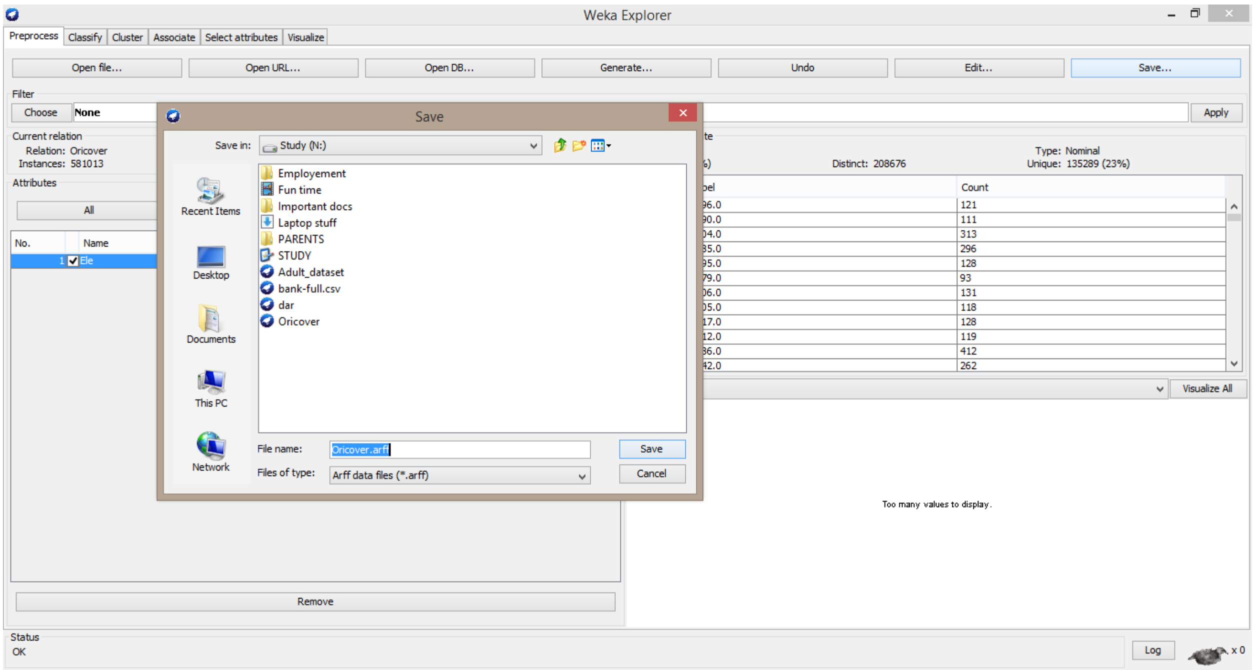
Employement
Fun time
Important docs
Laptop stuff
PARENTS
STUDY
bank-full
Oricover

File name: Oricover.csv
Files of type: CSV data files (*.csv)

many values to display.

OK

- Select one or more attributes (here Numeric only) in the left box and Click on Save.



Relation: Oricover
Instances: 581013

1 ☑ Ele

Save in: Study (N:)

File name: Oricover.arff

Files of type: Arff data files (*.arff)

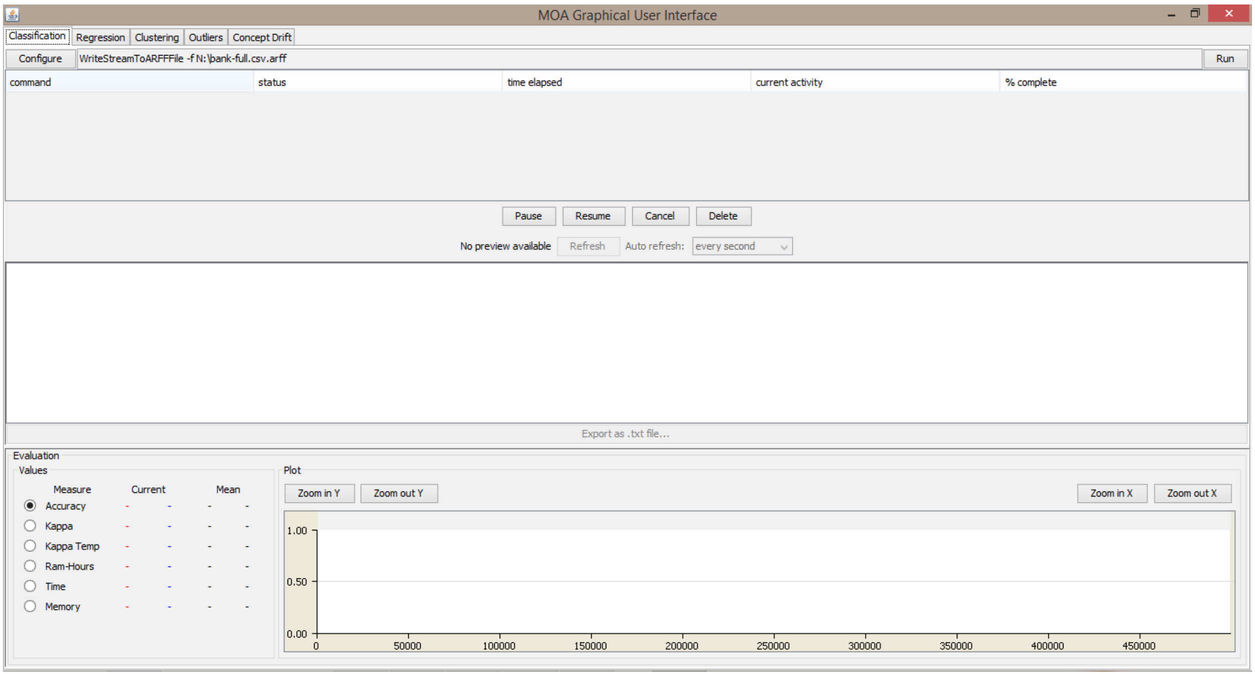
Distinct: 208676

Type: Nominal
Unique: 135289 (23%)

...el	Count
96.0	121
90.0	111
04.0	313
35.0	296
95.0	128
79.0	93
06.0	131
05.0	118
17.0	128
12.0	119
36.0	412
42.0	262

MOA (Massive Online Analysis)

MOA is used for performing data mining on data stream i.e. real time data like air traffic control, stock exchanges, etc. Current framework does not deal with real time data.



WriteStreamToARFFFile -f N:\bank-full.csv.arff

MOA is also an open platform which can be easily available on internet. It is a collection of various algorithms used perform data mining. “It implements boosting, bagging, and Hoeffding Trees, all with and without Naïve Bayes classifiers at the leaves. MOA supports bi-directional interaction with WEKA, the Waikato Environment for Knowledge Analysis, and is released under the GNU GPL license” [10]

Chapter 6
Inference from experimental study

For the experiment purpose, two datasets “Adult” and “Bank” are used. As discussed earlier this algorithm has some limitations like, it can be applied only on numeric dataset and only on one attribute at a time.

Below shown tables demonstrate the results acquired by applying the Geometric Data Perturbation algorithm on two numeric attributes of each data set.

As mentioned earlier in the implementation steps, perturbed dataset D’ is obtained by adding Gaussian Noise in the original dataset D and different factors of both the datasets are compared after applying two different algorithms on the original as well as perturbed datasets.

Here, WEKA tool is used to apply NB (Naïve Bayesian classifier algorithm) and J48 (Decision tree analysis algorithms). Below tables show the comparison between original values and modified values and their efficiencies.

Dataset 1: Adult dataset

	Age				Education			
	NB		J48		NB		J48	
	Original	Perturbed	Original	Perturbed	Original	Perturbed	Original	Perturbed
Correctly classified instances	0.8342	0.8318	0.8621	0.8573	0.8342	0.8291	0.8621	0.8562
Incorrectly classified instances	0.1657	0.1681	0.1378	0.1426	0.1657	0.1708	0.1378	0.1437
Time taken	0.2	0.23	4.52	4.18	0.2	0.22	4.84	5.04
Kappa Statistics	0.4993	0.4905	0.6004	0.5805	0.4993	0.475	0.6004	0.5721
Mean Absolute error	0.1735	0.1759	0.1942	0.2009	0.1735	0.1771	0.1942	0.2031
Root Mean Squared error	0.3723	0.3756	0.3196	0.3246	0.3723	0.3152	0.3196	0.3297
Relative absolute error	0.4745	0.4809	0.5309	0.5495	0.4745	0.4844	0.5309	0.5553
Root relative squared error	0.8706	0.8783	0.7474	0.7592	0.8706	0.8772	0.7474	0.7711

(Table 7.1 Adult Dataset)

From the table 7.1 we can see that after Applying NB (Naïve Bayesian) algorithm on both original and perturbed data, accuracy is nearly same in both scenarios.

NB	Original	Perturbed
Correctly Classified	0.8342	0.8318
Incorrectly Classified	0.1657	0.1681

This means with NB (Naïve Bayesian) algorithm on **original adult data** the accuracy of

Correctly classified instances was 83.42%

Incorrectly classified instances was 16.57%

This means with NB (Naïve Bayesian) algorithm on **perturbed adult data** the accuracy of

Correctly classified instances was 83.18%

Incorrectly classified instances was 16.81%

For this dataset, it proves that using Geometric data perturbation on dataset. The privacy of the original data can be preserved by little accuracy loss.

Similarly, after applying J48 on the original adult dataset:

NB	Original	Perturbed
Correctly Classified	0.8621	0.8573
Incorrectly Classified	0.1378	0.1426

This means with J48 algorithm on **original adult data** the accuracy of

Correctly classified instances was 86.21%

Incorrectly classified instances was 13.78%

This means with J48 algorithm on **perturbed adult data** the accuracy of

Correctly classified instances was 85.73%

Incorrectly classified instances was 14.26%

Results from both algorithms show that Geometric data perturbation can be a good alternative where accuracy is not utmost important for a data owner than the security of the data.

Dataset 2: Bank dataset

	Age				Duration			
	NB		J48		NB		J48	
	Original	Perturbed	Original	Perturbed	Original	Perturbed	Original	Perturbed
Correctly classified instances	0.8807	0.8805	0.9031	0.903	0.88	0.866	0.9031	0.8924
Incorrectly classified instances	0.1193	0.1195	0.0968	0.0969	0.1193	0.1339	0.0968	0.1075
Time taken	0.43	0.45	6.5	7.17	0.44	0.46	7.72	7.94
Kappa Statistics	0.4391	0.4346	0.4839	0.4846	0.4391	0.3413	0.4839	0.3354
Mean Absolute error	0.1532	0.1542	0.1269	0.1276	0.1532	0.1681	0.1269	0.157
Root Mean Squared error	0.3088	0.3075	0.2773	0.2781	0.3088	0.3305	0.2773	0.2986
Relative absolute error	0.7416	0.7464	0.6142	0.6175	0.7416	0.8135	0.6142	0.7596
Root relative squared error	0.9606	0.9567	0.8628	0.8653	0.9606	1.028	0.8628	0.9289

(Table 7.2 Bank Dataset)

- Flow of the Geometric data perturbation:

a. Geometric data perturbation technique has not been adapted for privacy preservation in data mining.
b. Loss of data is the main issue while performing data mining.
c. Issues with privacy on data stream mining.
d. With the change in distance relationship, accuracy of the classification model is reduced.

Chapter 7
Conclusion and Future Work

Geometric Transformation technique based on Multiplicative Data Perturbation approach has been applied for adding random noise to the original dataset to preserve privacy of sensitive attributes. This approach has been in direction to keep statistical relationship intact to mine useful results with perturbed data. It takes sensitive attributes as dependent attributes whereas, remaining attributes of dataset except class attribute are considered as independent attributes. Any calculations for adding tuple specific random noise is done only on dependent attributes of the dataset. The above framework uses Naïve Bayesian Classification algorithm to estimate the correct value of classification results from perturbed dataset over results from standard dataset. Accuracy of the results from the perturbed data will be less than the accuracy of the results from the original dataset. But, it is possible to achieve main objective of preserving the privacy of the sensitive info with less accuracy loss and the loss can further be minimized.

The paper uses Geometric Data Perturbation technique for numeric data only. The algorithm can be applied on non- numeric dataset using k-anonymization techniques. The algorithm can also be expanded to check real time data using stream analysis tool. The applied algorithm is working to extract single column value only and can be extended for more than one column at a time and also this algorithm can be applied to more classification as well as clustering algorithms. One of the future goals can also be to improve efficiency of the data mining of altered dataset and make privacy preserving more effective with minimal accuracy loss.

References

1. Osmar R Zalane, "Principles of Knowledge Discovery in Databases", CMPUT690, 1999
2. Oracle Data Mining Concepts, 11*g* Release 1 (11.1), May 2008[Online]. Available: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm
3. Martin Brown, "Data mining techniques", December 2012[Online]. Available: http://www.ibm.com/developerworks/library/ba-data-mining-techniques/
4. Bharati M. Ramageri, Lecturer, Dept. of Computer Application, MIITR, Pune-Maharashtra, India, "Data mining techniques and applications", Indian Journal of Computer Science and Engineering, Vol 1 No 4 301-305
5. 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. [Available Online: http://www.cs.waikato.ac.nz/ml/weka/index.html]
6. [Available Online: http://moa.cms.waikato.ac.nz/]
7. Vittorio Castelli, Research staff member, IBM T.J. Watson Research Center,[Available Online: http://www.ee.columbia.edu/~vittorio/Lecture-2.pdf]
8. Keke Chen [Dept. of CSE, Wright state university, Dayton], Ling Liu [CCGIT,Atlanta, GA], "Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining", [Online Available: http://knoesis.wright.edu/library/download/geometric_perturbation.pdf]
9. Twinkle Ankleshwaria, Prof. J. S. Dhobi,"Geometric Data Perturbation Approach For Privacy Preserving in data Stream Mining" Engineering Universe for Scientific research and Management, Impact factor 3.7, Volume 6, Issue 4, April 2014.
10. Hitesh Chhikaniwala [Ganpat University, India], Dr. Sanjay Garg [Nirma University, India],"Privacy Preserving Data Mining Techniques: Challenges and Issues"
11. Kun Liu, Hillol Kargupta, Senior Member, IEEE and Jessica Ryan, "Random Projection-based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE transaction on knowledge, Vol. 18, No. 1, January 2006
12. Stanley R. M. Oliveria, Osmar R. Zaiane, "Data Perturbation by rotation for privacy preserving Clustering", Technical Report TR 04-17, August 2004.
13. Charu C. Aggarwal and Philip S. Yu, "A Condensation Approach to Privacy Preserving Data Mining", IBM T. J. Watson Research Center, Hawthorne, NY
14. Prashant Lahane, R K Bedi, Prasad Halgaonkar, "Data Stream Mining", International Journal of Advances in computing and Information Researches, ISSN:2277-4068, Volume 1-No. 1, January 2012.
15. Ms. Ompriya Kale, Ms. Prachi Patel, "A Survey on Privacy Preserving Data Mining Techniques", Global journal of Advanced Engineering Technologies, ISSN: 2277-6370, vol2, Issue3-2013.

16. Neha Gupta, Indrajeet Rajput, ”Preserving privacy using data perturbation in Data Stream” International Journal of advanced Research in computer engineering & technology (IJARCET) Volume 2, No. 5, May 2013
17. Yabo Xu, Ke Wang, Ada Wai-Chee Fu,Rong She, and Jian Pei, Privacy-Preserving Data Stream Classification, springer,pp.489-510(2008).
18. Ching-Ming Chao, Po- Zung Chen and Chu- Hao Sun, Privacy-Preserving Classification of Data Streams, Tamkang Journal of Science and Engineering, Vol. 12, No. 3, pp. 321-330(2009).
19. R.VidyaBanu and N.Nagaveni,” Preservation of Data Privacy using PCA based Transformation”,in 2009 International Conference on Advances in Recent Technologies in Communicationand Computing, in 2009 IEEE computer society,p.439.
20. Kun Liu, HillolKargupta, Senior Member, IEEE, and Jessica Ryan, “Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 1, JANUARY 2006, p.92.
21. P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” Technical Report SRI-CSL-98-04, 1998.

Appendix A

➔ Java code to add the Gaussian Noise into the original dataset to get perturbed dataset

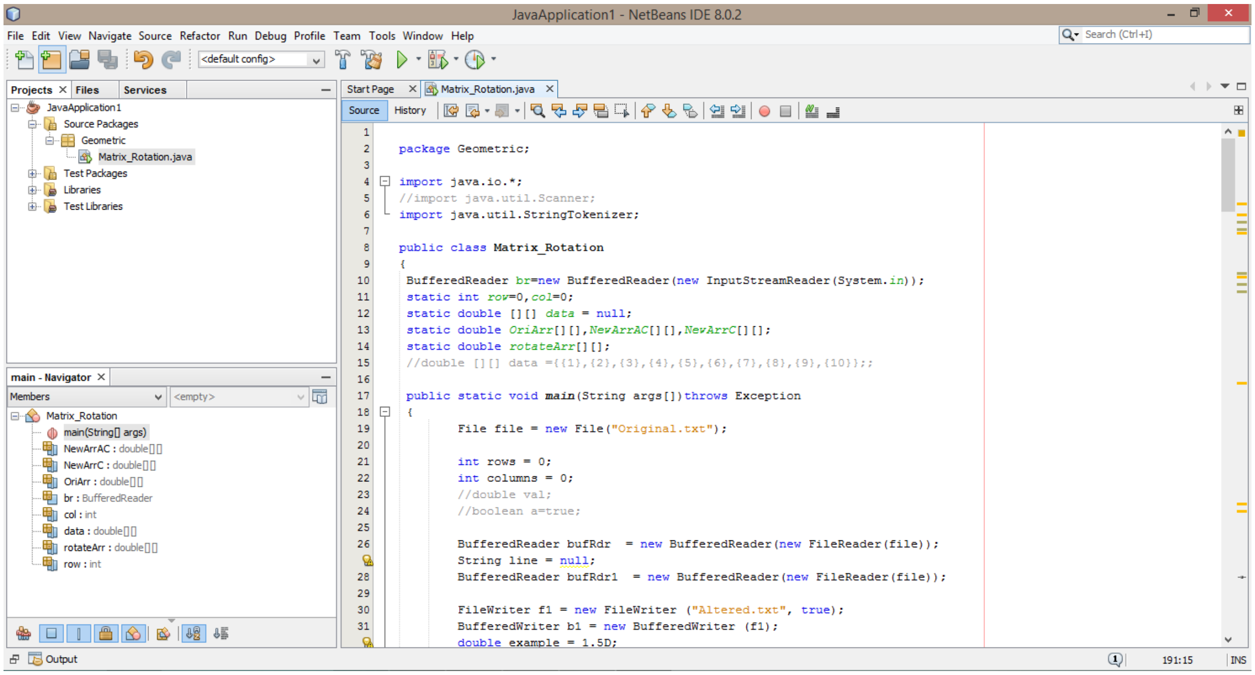
```
package Geometric;

import java.io.*;
import java.util.StringTokenizer;

public class Matrix_Rotation
{
 BufferedReader br=new BufferedReader(new InputStreamReader(System.in));
 static int row=0,col=0;
 static double [][] data = null;
 static double OriArr[][],NewArrAC[][],NewArrC[][];
 static double rotateArr[][];

 public static void main(String args[])throws Exception
 {
        File file = new File("Original.txt");

        int rows = 0;
        int columns = 0;
        BufferedReader bufRdr  = new BufferedReader(new FileReader(file));
        String line = null;
        BufferedReader bufRdr1  = new BufferedReader(new FileReader(file));

        FileWriter f1 = new FileWriter ("Altered.txt", true);
        BufferedWriter b1 = new BufferedWriter (f1);
        double example = 1.5D;

        int q1;
        int q2;
        q1=rows;
        q2=columns;
```

```
  while((line = bufRdr.readLine()) != null )
  {
            rows++;
  }
  columns++;
  b1.newLine();
  double [][] dataset = new double [rows][columns];

  while((line = bufRdr1.readLine()) != null )
  {

     StringTokenizer st = new StringTokenizer(line,",");

     while (st.hasMoreTokens())
  {
     //get next token and store it in the array
      dataset[q1][q2]=Double.parseDouble(st.nextToken());
      example=dataset[q1][q2];
      example=example*1.0D;
  }
     System.out.println(dataset[q1][q2]);
     q1++;
  }
   q2++;
System.out.println("");
System.out.println("row:"+rows);
System.out.println("column:"+columns);
OriArr=new double[rows][columns];
NewArrC=new double[rows][columns];
for(int i=0;i<rows;i++)
{
     System.arraycopy(dataset[i], 0, OriArr[i], 0, columns);
}
System.out.println("\nORIGINAL:-");
for(int i=0;i<rows;i++)
{
   for(int j=0;j<columns;j++)
   {
    NewArrC[i][j] = OriArr[i][j];
     System.out.println(NewArrC[i][j]);
```

```
         }
     }
     System.out.println("\nROTATED:-");
int x=columns-1;
int y1=rows-1;
int k1=0;
for(int i=y1;i>=0;i--)
{
    NewArrC[k1][columns-1]=OriArr[i][columns-1];
    k1++;
}
           for(int i=0;i<rows;i++)
           {
             double example1=NewArrC[i][0];
             example1=example1*1.0D;
             System.out.println(NewArrC[i][0]);
    }
double ans1[][]=new double [rows][columns];
System.out.println("\nMULTIPLICATION:-");
for(int y=0;y<rows;y++)
{
    ans1[y][0]=OriArr[y][0]*NewArrC[y][0];
    double example2=ans1[y][0];
             example2=example2*1.0D;
    System.out.println(example2);
}

double example21;
double sum=0;
double mean;
double TransArr[][]=new double [rows][columns];
     for (double[] OriArr1 : OriArr) {
         sum = sum + OriArr1[0];
     }
mean=sum/OriArr.length;
example21=mean*1.0D;
System.out.println("\nMEAN VALUE:-");
System.out.println(example21);

System.out.println("\nTRANSLATION:-");
```

```
for(int hg=0;hg<OriArr.length;hg++)
{
   TransArr[hg][0]=OriArr[hg][0]+mean;
   example21=TransArr[hg][0]*1.0D;
   System.out.println(example21);
}

// code for gaussien noise
double Mean;
double sum1=0.0;
    for (double[] OriArr1 : OriArr) {
        sum1 = sum1 + OriArr1[0];
    }
Mean=sum1/OriArr.length;

double answer1 = 0;
double answer2 = 0;
double answer3=0;
    for (double[] OriArr1 : OriArr) {
        answer1 = OriArr1[0] - Mean;
        answer1=answer1*answer1;
        answer2=answer2+answer1;
    }
answer3=answer2/OriArr.length;
double varience1=Math.sqrt(answer3);
double ansq=2*Math.PI;
double ansd=Math.sqrt(ansq);
double f11=varience1*ansd;
double term1=1/f11;
double random=OriArr[4][0];
double rx=random-Mean;
rx=rx*rx;
double df=answer3*2;
double hj=rx/df;
double term2=Math.pow (Math.E, hj);
double pdf=term1*term2;
System.out.println("\nGAUSSIAN NOISE:-");
System.out.println(pdf);
double finalGx[][]=new double[rows][columns];
 b1.write("GEOMETRIC DATA PERTURBATION:-");
 b1.newLine();
System.out.println("\nFINAL RESULT:-");
for(int yh=0;yh<OriArr.length;yh++)
{
   finalGx[yh][0]=ans1[yh][0]+TransArr[yh][0]+pdf;
   System.out.println(finalGx[yh][0]);
   b1.write(String.valueOf(finalGx[yh][0]));
   b1.newLine();
}
b1.close();
 }
}
```

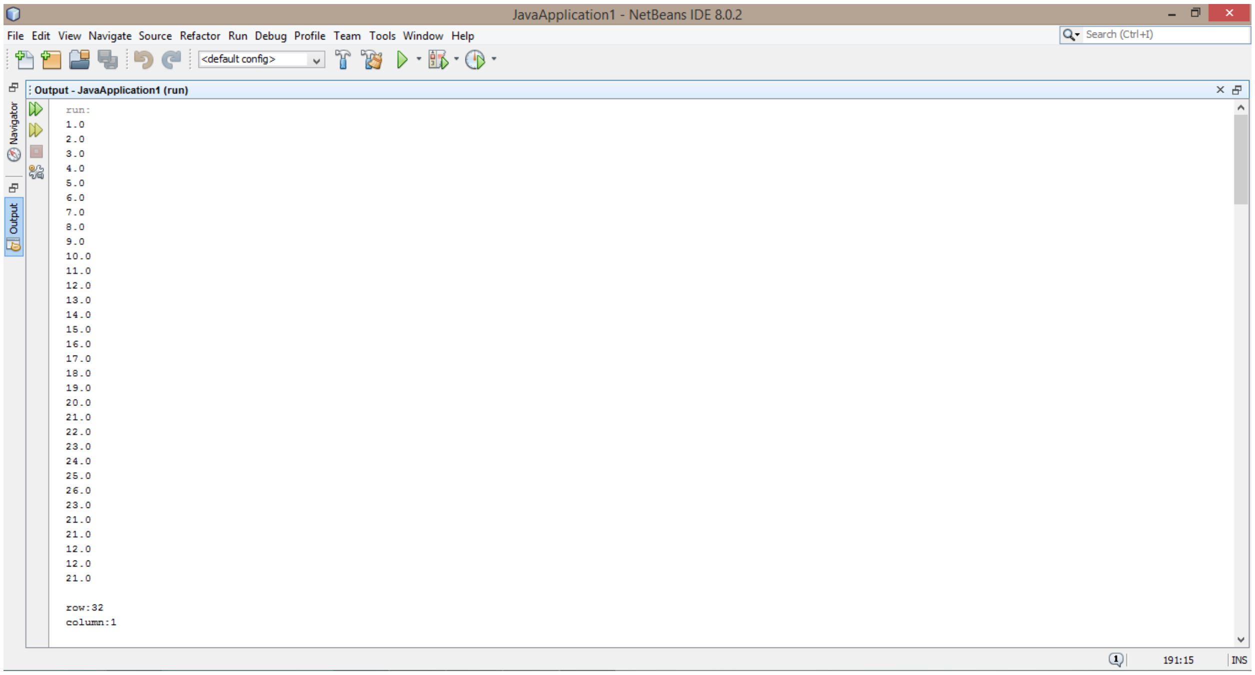
➔ NetBeans IDE 8.0.2



```

package Geometric;

import java.io.*;
//import java.util.Scanner;
import java.util.StringTokenizer;

public class Matrix_Rotation
{
 BufferedReader br=new BufferedReader(new InputStreamReader(System.in));
 static int row=0,col=0;
 static double [][] data = null;
 static double OriArr[][],NewArrAC[][],NewArrC[][];
 static double rotateArr[][];
 //double [][] data ={{1},{2},{3},{4},{5},{6},{7},{8},{9},{10}};;

 public static void main(String args[])throws Exception
 {
        File file = new File("Original.txt");

        int rows = 0;
        int columns = 0;
        //double val;
        //boolean a=true;

        BufferedReader bufRdr  = new BufferedReader(new FileReader(file));
        String line = null;
        BufferedReader bufRdr1  = new BufferedReader(new FileReader(file));

        FileWriter f1 = new FileWriter ("Altered.txt", true);
        BufferedWriter b1 = new BufferedWriter (f1);
        double example = 1.5D;
```

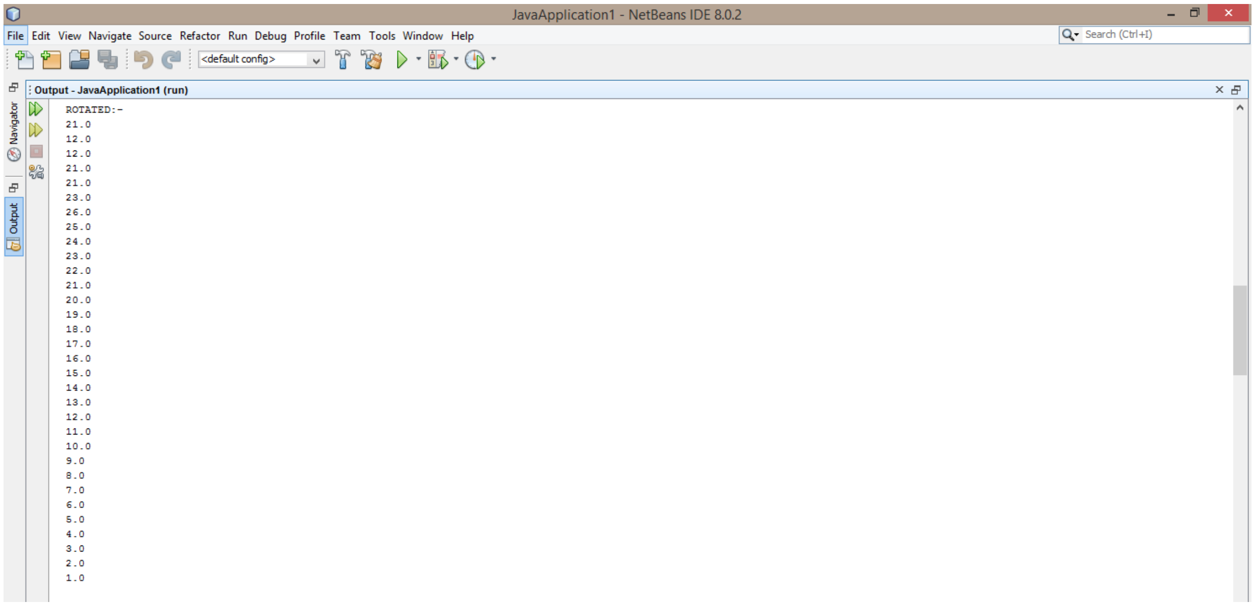
➔ Steps to get the desired output.

1. Read the original data from Text file. (here, Original.txt) and convert it to Decimal


JavaApplication1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Search (Ctrl+I)
<default config>
Output - JavaApplication1 (run)
Navigator
Output
191:15
INS

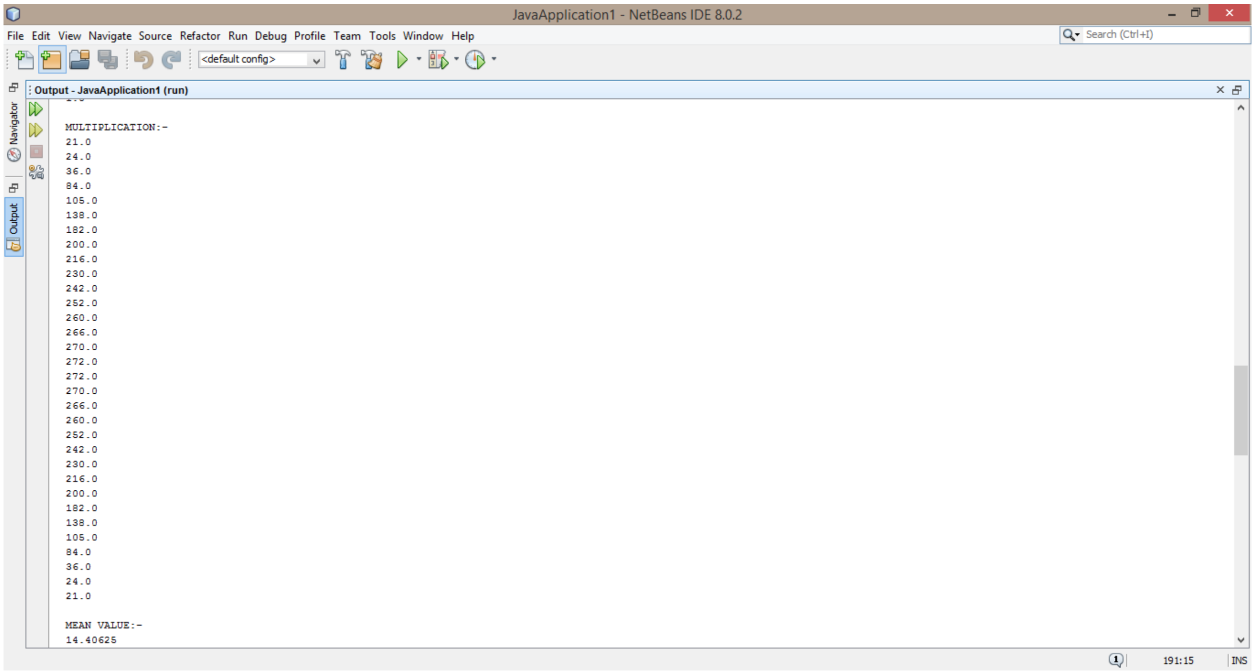
```
run:
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
23.0
21.0
21.0
12.0
12.0
21.0

row:32
column:1
```

2. Now rotate the data.


JavaApplication1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Search (Ctrl+I)
<default config>
Output - JavaApplication1 (run)
Navigator
Output

```
ROTATED:-
21.0
12.0
12.0
21.0
21.0
23.0
26.0
25.0
24.0
23.0
22.0
21.0
20.0
19.0
18.0
17.0
16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
```

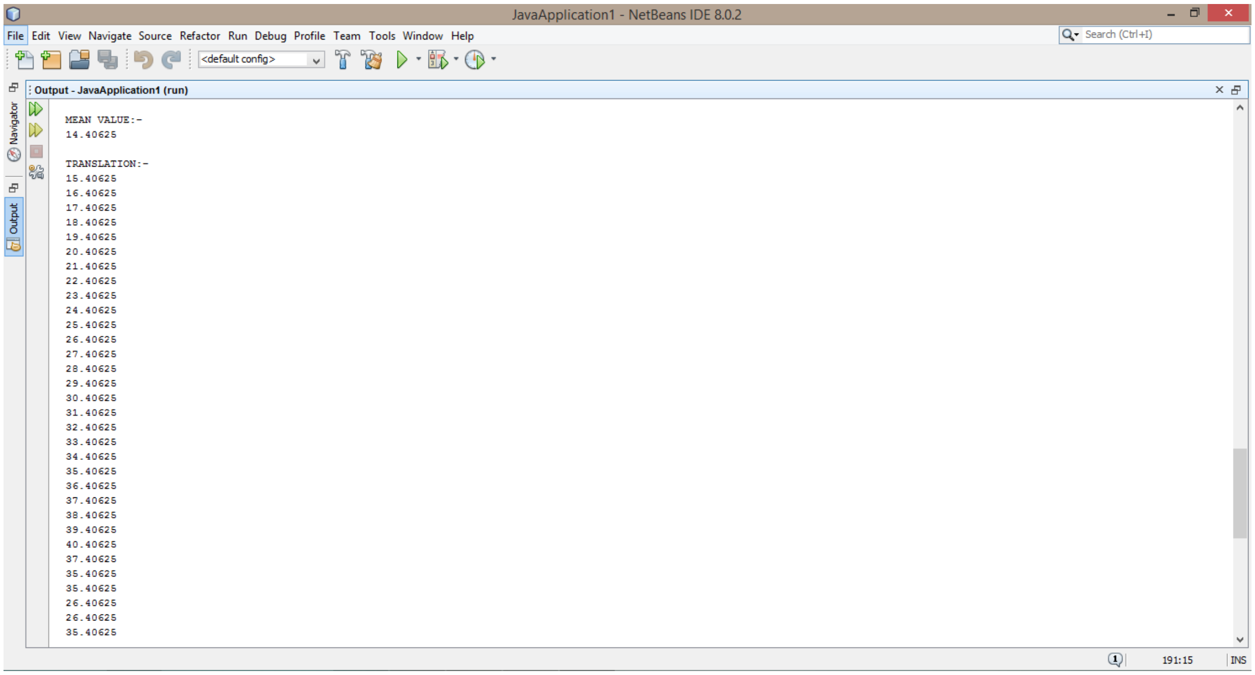
3. Perform multiplication of original and rotated data.



JavaApplication1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Search (Ctrl+I)
<default config>
Navigator
Output
Output - JavaApplication1 (run)
191:15
INS

```
MULTIPLICATION:-
21.0
24.0
36.0
84.0
105.0
138.0
182.0
200.0
216.0
230.0
242.0
252.0
260.0
266.0
270.0
272.0
272.0
270.0
266.0
260.0
252.0
242.0
230.0
216.0
200.0
182.0
138.0
105.0
84.0
36.0
24.0
21.0

MEAN VALUE:-
14.40625
```

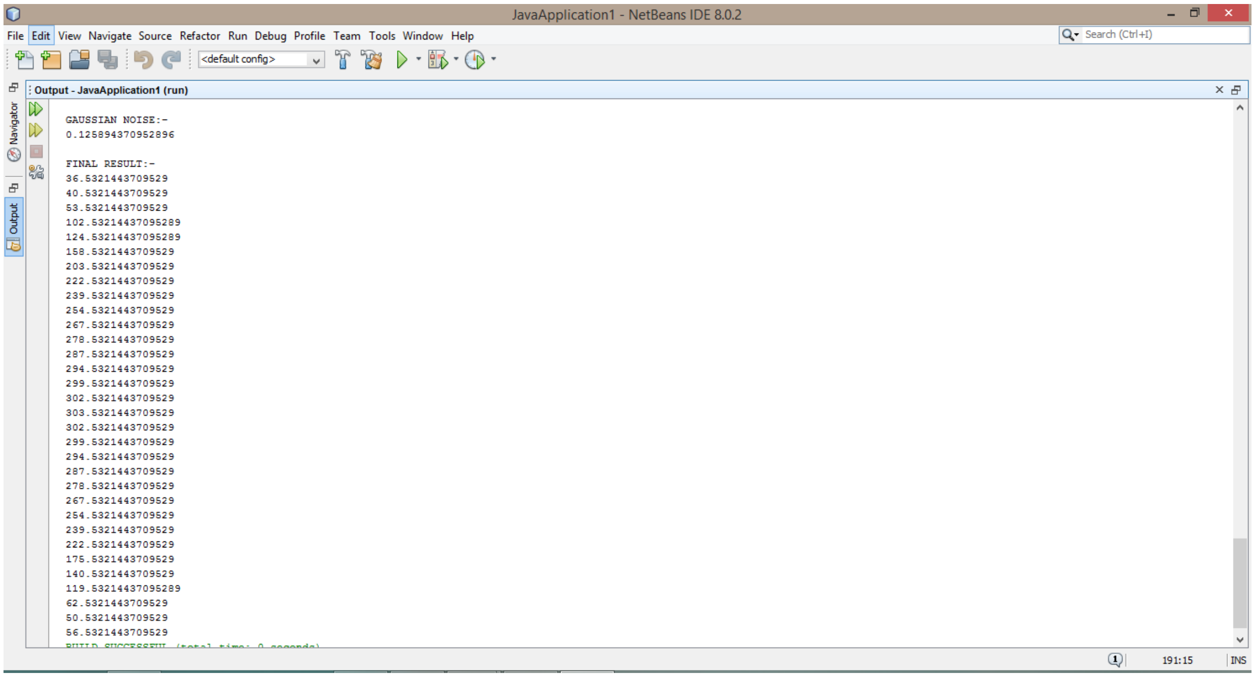
4. Find the Mean value and perform translation



JavaApplication1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Search (Ctrl+I)
<default config>
Navigator
Output
Output - JavaApplication1 (run)
191:15
INS

```
MEAN VALUE:-
14.40625

TRANSLATION:-
15.40625
16.40625
17.40625
18.40625
19.40625
20.40625
21.40625
22.40625
23.40625
24.40625
25.40625
26.40625
27.40625
28.40625
29.40625
30.40625
31.40625
32.40625
33.40625
34.40625
35.40625
36.40625
37.40625
38.40625
39.40625
40.40625
37.40625
35.40625
35.40625
26.40625
26.40625
35.40625
```

5. Find the Gaussian Noise value and print the Final result


JavaApplication1 - NetBeans IDE 8.0.2
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
Search (Ctrl+I)
<default config>
Navigator
Output
Output - JavaApplication1 (run)
191:15
INS

```
GAUSSIAN NOISE:-
0.125894370952896

FINAL RESULT:-
36.5321443709529
40.5321443709529
53.5321443709529
102.53214437095289
124.53214437095289
158.5321443709529
203.5321443709529
222.5321443709529
239.5321443709529
254.5321443709529
267.5321443709529
278.5321443709529
287.5321443709529
294.5321443709529
299.5321443709529
302.5321443709529
303.5321443709529
302.5321443709529
299.5321443709529
294.5321443709529
287.5321443709529
278.5321443709529
267.5321443709529
254.5321443709529
239.5321443709529
222.5321443709529
175.5321443709529
140.5321443709529
119.53214437095289
62.5321443709529
50.5321443709529
56.5321443709529
BUILD SUCCESSFUL (total time: 0 seconds)
```

This output is automatically stored into the Text file. (Here, Altered.txt)