ISSN 2319-2720
Volume 4, No.1, January – March 2015
International Journal of Computing, Communications and Networking
Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn01412015.pdf

Preserving Privacy in Data Mining by Data Perturbation Technique

Obulapu Hitesh Reddy[1], **Pardeep Singh**[2]
[1]NIT Hamirpur, Himachal Pradesh, India, hiteshreddy1222@gmail.com
[2]NIT Hamirpur, Himachal Pradesh, India, avagaman@gmail.com

ABSTRACT

Data mining is a methodology which is used for extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large database. The general aim of the data mining process is to extract statistics from a data set and change it into a reasonable construction for advance procedure. Data mining has a number of applications, like medical, business, scientific, human life and education etc.

During the process of data mining, the data sets are accessed by a number of process/modules for the extraction of data. This may lead to disclosure of sensitive information and hence a breach of privacy.

The major challenge of data perturbation is to stabilize privacy protection and data quality. Perturbation of data is to accomplish by anticipating outcome among the level of data confidentiality and the level of data value. Recently, several techniques in data mining for preserving privacy has been proposed. The current research technique used for privacy preserving data mining is Hybrid Approach, which uses, A combination of k-Anonymity and Randomization approaches which have better accuracy and also facilitates the reconstruction of the original data.

In this paper, we concentrated on data perturbation procedures, i.e., Adding noise to the data in command to check thorough release of trusted values. The additive noise still permits the aggregate information to be read, about the overall collection of data but does not give away accurate values. The noise is a small randomly generated (or using certain algorithms), and added to the data. Hence, by this method we protect individual information and release information at the same time.

Key words: data mining, data perturbation, privacy preserving techniques, randomization

1. INTRODUCTION

Data mining is the procedure of seeing fascinating facts from enormous volumes of data stockpiled either in databases, data warehouses, or other data repositories. Data mining plays a vital role as a frontier in database systems and one of the most promising interdisciplinary growth in the information industry [1]. Due to the advances in information processing technology and the storage capacity, modern organization collects a large amount of data [2]. The huge amount of specific personal information is daily collected and analyzed by the various applications with the help of data mining.

For extracting hidden and earlier unknown information from such enormous data sets the organizations rely on several data mining techniques. Throughout the entire procedure of data mining these information frequently gets visible to various events. So delicate material kept round the single can actually be revealed resulting in a breach of individual confidentiality. So, we require skills for defensive individual secrecy while agreeing to data mining.

Several data mining privacy preserving techniques has been proposed, in order to protect the sensitive data information stored in the volumes. Public cognizance of privacy and the absence of public faith in the administration may publicize more difficulty to data collection. Therefore, appropriate privacy preserving techniques has a major role in data mining [2].

Privacy preserving data mining has been intensively studied since 2000. Data perturbation technique, first proposed by Y. Lindell and B. Pinkas [7], represents a cryptographic technique through which sensitive data can be encrypted. In this paper, we concentrated on data perturbation procedures. This procedure is generally used in situations where individuals reporting data to a data miner, can be ruffle the exact data with some kind of recognized random noise and report the noisy information to the data miner [3].

The requirement is for an additional layer of software between the database and the user that takes care of perturbation of the data. The idea of the paper is to develop Java modules, which perform perturbation of the data before the user accesses it from the database. The proposed application enables the export of perturbed data.

1.1 ORGANIZATION OF PAPER

In this paper, we discuss various methods and techniques in the area of privacy preserving data mining. The rest of the paper is organized as follows. In section 1, we give the basic idea about data mining, privacy and problem statement. In section 2, in literature survey, we discuss some previous research works on data perturbation techniques. In section 3, methodology, we see about requirements of software and database connectivity. In section 4, existing and proposed masking techniques is discussed. In section 5, discusses results and performance of WEKA tool analysis. And finally we conclude the paper in section 6.

2. LITERATURE SURVEY

In this section we discuss few present methods for privacy preserving data mining deals with securing the privacy of specific data or delicate information without losing the effectiveness of the data [1]. Data perturbation techniques are regularly used to secure top secret information from unauthorized requests while providing supreme access and exact data to legitimate queries [4].

A number of different techniques have been proposed for privacy preserving data mining. Privacy preserving techniques can be categorized based upon data distribution, data type, data mining tasks and protection methods. Confidentiality can be secure over and done with various approaches such as data variation and confident multi party computation. Privacy preserving performances can also be ordered based upon security methods such as Suppression, Data swapping, Aggregation and Noise addition techniques. Masking methods can function on different data types. Data types can be classified into continuous variables and categorical variables. Masking methods are categorized into perturbative and non perturbative masking techniques. Perturbative method is nothing but modifying the original data, by using various techniques like, Micro aggregation, Additive noise, Rank swapping, Randomization, Rounding and Resampling etc. While some records are repressed and/or some information are detached, In non pertubative technique, but original data cannot be modified [4].

Data perturbation technique, first proposed by Y. Lindell and B. Pinkas, signifies a cryptographic procedure complete which delicate data can be encoded and the outcome is scaled when more than a small number of events are involved. Data perturbation contains a different type of techniques like clustering techniques, additive perturbation, multiplicative and randomized data perturbation.

2.1 Existing Clustering Techniques

Major clustering approaches are hierarchical, partitioning algorithms, density-based, grid based and model based. Density based clustering is based on local cluster criteria, such as density connected points. Clustering techniques used in perturbation are CACTUS, ROCK, and COOLCAT.

2.1.1 Ganti proposed [CACTUS] in 1999

Definite Grouping uses outlines (CACTUS), is a fast summarization based algorithm which uses swift data to catch healthy penalized groups. The Central idea of CACTUS is data summary (inter and intra attribute summary) is sufficient enough to find candidate cluster which can then be validated. Unique sets of bunches are exclusively recognized by a central set of characteristic values that take place in no other bunch. The distinguishing sets that attribute value sets that uniquely occur within only one cluster. CACTUS has three phases clustering algorithm, summarization phase which, computes the summary information, clustering phase, which discover a set of candidate clusters and validation phase which determine the actual set of clusters [6].

2.1.2 Guha proposed [ROCK] in 1999 is Robust clustering using links (ROCK) is an agglomerative hierarchical clustering algorithm in the text of the market basket dataset. It uses links to measure similarity proximity and cubic computational complexity by (1). Rock associates the total of mutual neighbors for the two topics. In the beginning, every tuple is allocated to a distinct group and then groups are combined frequently according to the nearness among groups. The nearness between groups is clear as the sum of the number of associates among all sets of tuples, everyplace the number of associates characterizes the number of mutual neighbors among two groups [5].

$$O(n^2+\mathrm{n}m_m m_a+n^2\log \mathrm{n}) \quad (1)$$

2.2 Additive Perturbation

The additive perturbation technique is masking the characteristic value by adding noise to the unique data. The procedure adds the noise to the data so that individual records should not be recovered, it will preserve privacy [8]. In late 80s to 90s, this method used in statistical databases to protect sensitive attributes. The additive perturbation method can generate the perturb data Z by adding the original value X with random noise Y this can be represented as follows (2). Information Z and the constraints of Y are available. The benefit of this method is they allow distribution reconstruction and permit individual user to do perturbation, they publish the noise distribution. The typical additive perturbation method is a column based additive randomization [9]. Column distribution based algorithms used Navie baye’s classifier and Decision tree methods.

$$Z = X+Y \quad (2)$$

2.3. Multiplicative perturbation

The Multiplicative perturbation is used to get good results for privacy preserving data mining. This method preserves the inter record distances roughly, and therefore the different records can be used in coincidence with the various distance intensive data mining applications [8]. Several of these methods originate their roots in the work of which shows how to use multi- dimensional projections in order reduce the dimensionality of the data [10]. Concentrated secrecy protective data mining can be done by Multiplicative perturbation. Multiplicative perturbations are classified into Geometric Data Perturbation (GDP) and Random projection perturbation (RPP). The Geometric Data Perturbation (GDP) consists of various techniques like Rotation data perturbation, translation data perturbation and noise addition. The Random projection perturbation method can generate the perturb data F(X) by matrix multiplication X is m * n matrix, (where m columns and n rows) and P is a k*m random matrix, k<=m. This can be represented as follows (3).

$$F(X) = P * X \tag{3}$$

2.4. Randomization technique

The randomization method is a procedure for confidentiality, maintaining data mining in which noise is more of the data in command to cover the characteristic standards of records [15]. The noise added is adequately huge, so that single record values cannot be recovered [11]. Therefore, the techniques are designed to drive aggregate distributions from the perturbed records. Next, data mining approaches can be advanced with instruction to work with these total deliveries [12]. This process has been usually used in the situation of misrepresenting data by possibility delivery for approaches such as surveys which have an indirect answer bias because of secrecy concerns [13], [14].

3. METHODOLOGY

In this section we discuss about the requirements required to develop the Java module and data mining tools used algorithms for identifying the most predictive attributes in the data.

3.1. JAVA - Integrated Development Environment, for programming

Eclipse is a multi-language software development environment comprising a base workspace and an extensible plug-in system for customizing the environment. It is written mostly in Java. Various languages are used in this software to create applications. Released under the terms of the Eclipse Public License, Eclipse SDK is free and open source software. The Eclipse SDK consist of the Eclipse Java development tools (JDT), proposing an IDE with a constructed-in incremental Java compiler and a complete model of the Java basis records. This permits for future refactoring methods and code analysis.

3.2. Database Management System

MySQL is the world's most used open source relational database management system (RDBMS) that runs as a server providing multiuser access to a number of databases. MySQL is a relational database management system (RDBMS), and transports through no GUI tools to run a MySQL database or be able to data contained in the databases. Users might use the contained within thorough knowledge line tools, or use MySQL "front ends", desktop software and web requests that generate and achieve MySQL databases, build database structures, back up data, look over status, and work with data archives. The authorized set of MySQL front-end tools, MySQL workbench is keenly advanced by oracle, and is at liberty offered for use.

3.3. Database Front-End

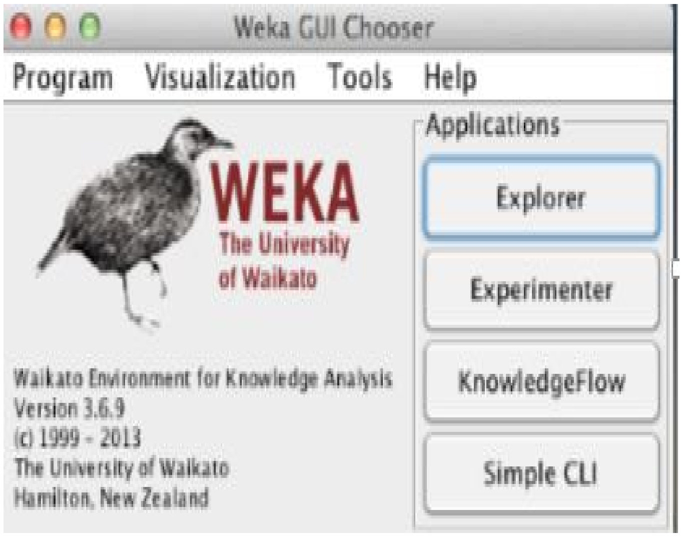
Sequel Pro is a fast, easy-to-use Mac database management application for working with MySQL databases. Sequel Pro provides you straight entrance to your MySQL database on local and remote servers.

3.4. Java Database Connectivity [JDBC]

JDBC is a Java oriented data entrance tools (Java standard edition platform) from Oracle Corporation. A client can access a database through an API, which helps in modifying and querying phenomena occurring in a database.

3.5. The Data Mining Tool

For the analysis of the statistics obtained on the data sets, we use a data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. It is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.



Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 1: Weka GUI Chooser

Weka is an open source software existing in the GNU General Public License. It is a cross platform and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Weka is also well suited for developing new machine learning schemes.

To demonstrate the similar results when mined, we use an example dataset and perform one of the functions/primitives of Weka. The select element panel provides algorithms for detecting the most predictive elements in a dataset. It uses any one of the following search algorithms to perform the operation like Best First, Greedy, Random, Ranker, Rank search, Exhaustive, Genetic and Linear Forward Selection. In this paper, we use a Best Frist algorithm to identify the most predictive attributes in the datasets

3.6. Block Diagram

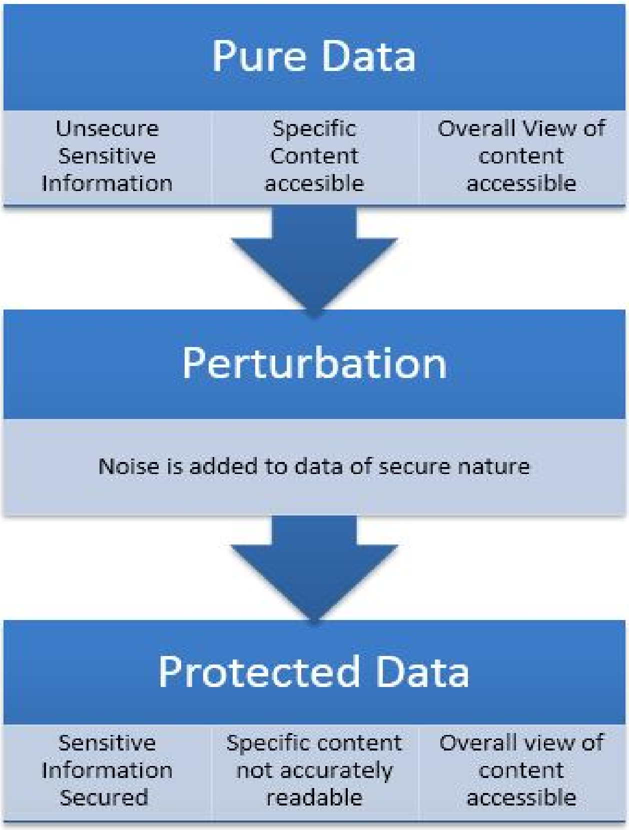

Pure Data
Unsecure Sensitive Information
Specific Content accesible
Overall View of content accessible
Perturbation
Noise is added to data of secure nature
Protected Data
Sensitive Information Secured
Specific content not accurately readable
Overall view of content accessible

Figure 2: Block diagram

3.7. Data Flow Diagram

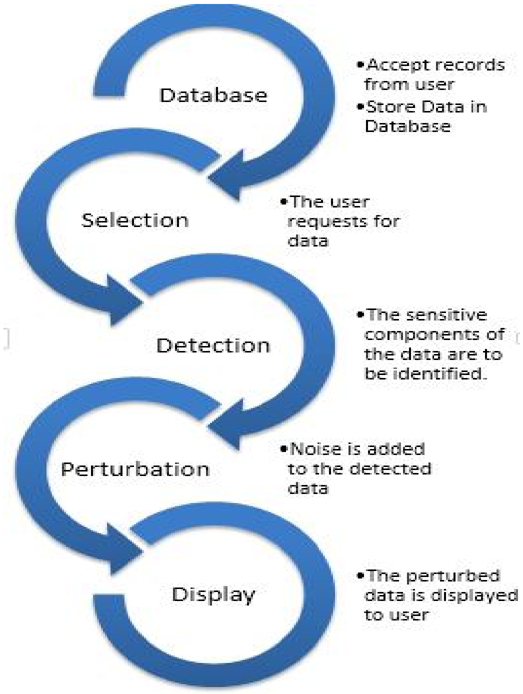

Database
•Accept records from user
•Store Data in Database
Selection
•The user requests for data
Detection
•The sensitive components of the data are to be identified.
Perturbation
•Noise is added to the detected data
Display
•The perturbed data is displayed to user

Figure 3: Flow chart

3.8. To Find the Sensitive Data (on which to apply perturbation)

```
Find Sensitivedata (string key words[], int priority[])
    {
        FinalScore <- 0
        ColumnNames <- field.ColumnNames()
        for i=1 to Key words.Length
        Do
          if ColumnNames.contains (Keywords[i])
          Do
            if FinalScore < Priority (i)
            Do
              FinalScore = priority(i)
              Field.SensitiveScore = finalscore
    }
```

3.9. Pertubation (noise addition)

```
Perturbation(int a[], int size)
{
    sar =a
    tot <- 0, avg <- 0, c1 <- 0, c2 <- 0
    for i=1 to sar.Length
    Do
      tot += sar(i)
      avg <- tot/sar.length
    for i=1 to sar.Length
    Do
        if avg <=sar(i)
        Do
          c1++
   Else
        Do
        c2++
        m1 <- 0
        result <- 0
     for i=1 to sar.Length
         Do
           if avg>sar[i]
             do
               m1 = 2*(avg/c1)
               if avg < sar[i]
               do
```

```
            m1 = 2*(avg/c2)
            result [i] <- m1
            Return result
    }
```

4. TESTING

In this section we see the user interface functions developed for import table, Detect Sensitive Fields, Perturbation and Export

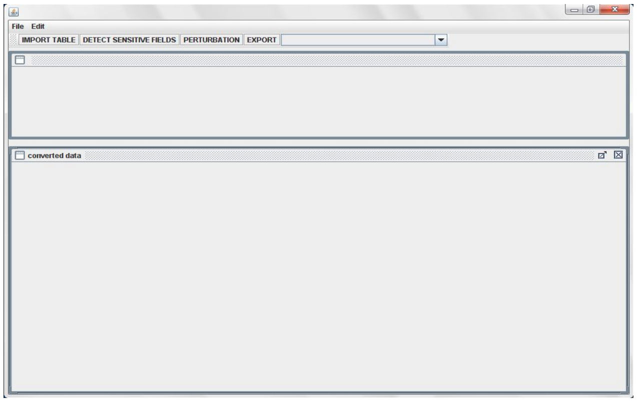
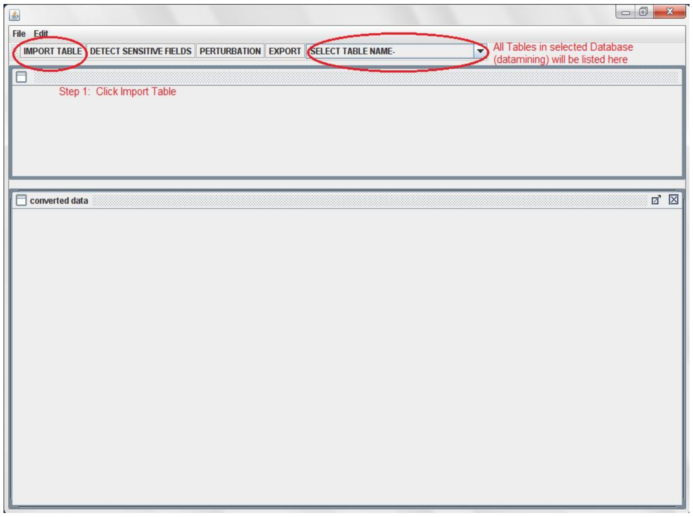


Figure 4: The Application window

The application window is a simple UI interface developed with Java's built in drawing tools (Applets) and consists of 5 functions [Import Table, Detect Sensitive Fields, Perturbation and Export] and 2 frames. One frame displays the imported records from the database, while the other will display the resulting table after the perturbation process.



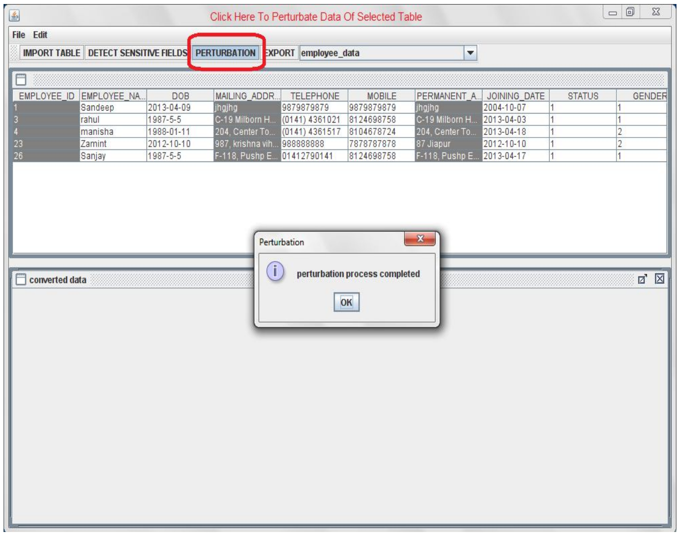
All Tables in selected Database (datamining) will be listed here

Step 1: Click Import Table

Figure 5: Import table

The import table function, loads the tables from a selected database into the application. Once, the tables are loaded, we can select one of the tables to work on.

The detecting sensitive fields are used to detect the unsecure sensitive information and overall view of the content accessible



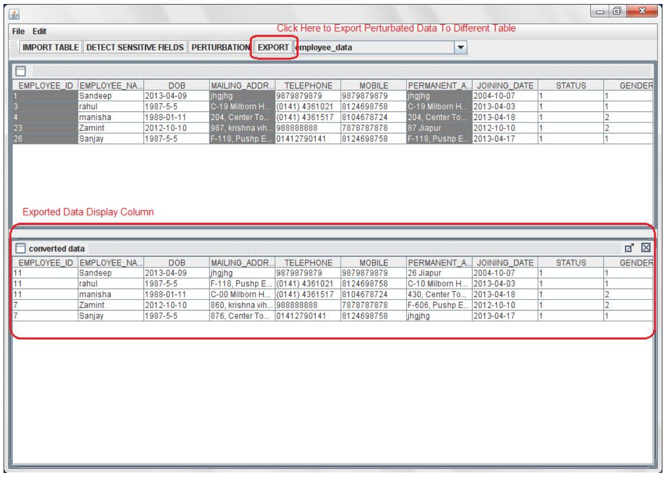
Click Here To Perturbate Data Of Selected Table

EMPLOYEE_ID	EMPLOYEE_NA...	DOB	MAILING_ADDR...	TELEPHONE	MOBILE	PERMANENT_A...	JOINING_DATE	STATUS	GENDER
1	Sandeep	2013-04-09	jhgjhg	9879879879	9879879879	jhgjhg	2004-10-07	1	1
3	rahul	1987-5-5	C-19 Milborn H...	(0141) 4361021	8124698758	C-19 Milborn H...	2013-04-03	1	1
4	manisha	1988-01-11	204, Center To...	(0141) 4361517	8104678724	204, Center To...	2013-04-18	1	2
23	Zamint	2012-10-10	987, krishna vih...	988888888	7878787878	87 Jiapur	2012-10-10	1	2
26	Sanjay	1987-5-5	F-118, Pushp E...	01412790141	8124698758	F-118, Pushp E...	2013-04-17	1	1

perturbation process completed

Figure 6: Perturbation

Then noise is added to the data of a secure nature in the perturbation and then display the perturbed data to the user. The perturbation technique performed by selecting the attribute values and altering an attribute value with a new value. After performing perturbation technique the sensitive data has been perturbed with new values. Later, export perturbed data to different values. The exported data display with new values.



Click Here to Export Perturbated Data To Different Table

EMPLOYEE_ID	EMPLOYEE_NA...	DOB	MAILING_ADDR...	TELEPHONE	MOBILE	PERMANENT_A...	JOINING_DATE	STATUS	GENDER
1	Sandeep	2013-04-09	jhgjhg	9879879879	9879879879	jhgjhg	2004-10-07	1	1
3	rahul	1987-5-5	C-19 Milborn H...	(0141) 4361021	8124698758	C-19 Milborn H...	2013-04-03	1	1
4	manisha	1988-01-11	204, Center To...	(0141) 4361517	8104678724	204, Center To...	2013-04-18	1	2
23	Zamint	2012-10-10	987, krishna vih...	988888888	7878787878	87 Jiapur	2012-10-10	1	2
26	Sanjay	1987-5-5	F-118, Pushp E...	01412790141	8124698758	F-118, Pushp E...	2013-04-17	1	1

Exported Data Display Column

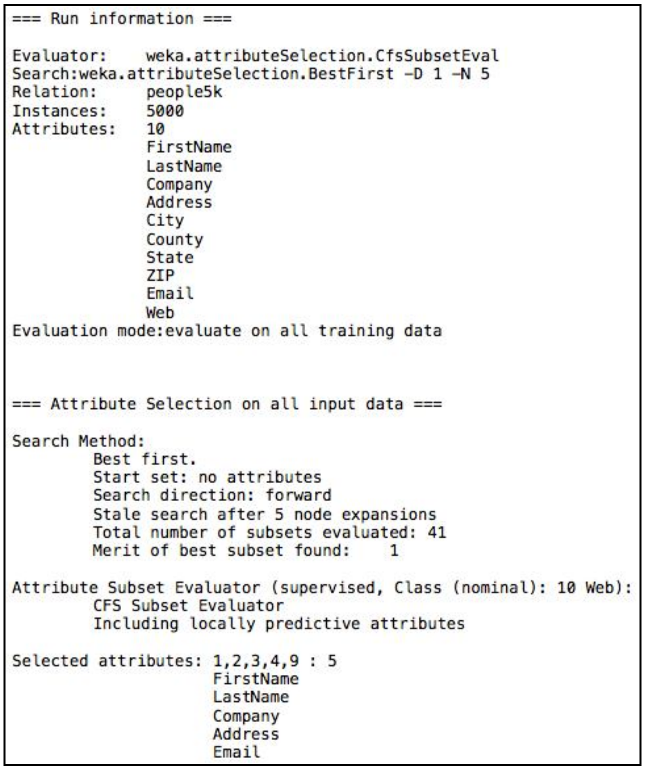
EMPLOYEE_ID	EMPLOYEE_NA...	DOB	MAILING_ADDR...	TELEPHONE	MOBILE	PERMANENT_A...	JOINING_DATE	STATUS	GENDER
11	Sandeep	2013-04-09	jhgjhg	9879879879	9879879879	26 Jiapur	2004-10-07	1	1
11	rahul	1987-5-5	F-118, Pushp E...	(0141) 4361021	8124698758	C-10 Milborn H...	2013-04-03	1	1
11	manisha	1988-01-11	C-00 Milborn H...	(0141) 4361517	8104678724	430, Center To...	2013-04-18	1	2
7	Zamint	2012-10-10	860, krishna vih...	988888888	7878787878	F-606, Pushp E...	2012-10-10	1	2
7	Sanjay	1987-5-5	876, Center To...	01412790141	8124698758	jhgjhg	2013-04-17	1	1

Figure 7: Export data

5. RESULTS AND DISCUSSIONS

The Perturbation is nothing but altering an attribute value with a new value. The data sets are distorted before publication. The provided data misleads in a way that disturbs the secured data set. With above methods the candidate dataset changed and produce a unique combination with more data items in the perturbed dataset; in perturbation technique data computed on the perturbed dataset do not vary from the data obtained on the unique data set.

5.1 Comparison of Pure Data and Perturbted Data



```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     people5k
Instances:    5000
Attributes:   10
              FirstName
              LastName
              Company
              Address
              City
              County
              State
              ZIP
              Email
              Web
Evaluation mode:evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 41
        Merit of best subset found:    1

Attribute Subset Evaluator (supervised, Class (nominal): 10 Web):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3,4,9 : 5
                     FirstName
                     LastName
                     Company
                     Address
                     Email
```

Figure 8: Evaluation of Pure data

To demonstrate the function of the application, we need a relevant dataset. A dataset that contains potential sensitive information, to be perturbed. We use sample data sets of 500, 5000 and 500000 records respectively. The table contains the following attributes/columns like first name, last name, company, address, city, country, state, zip and E-mail.

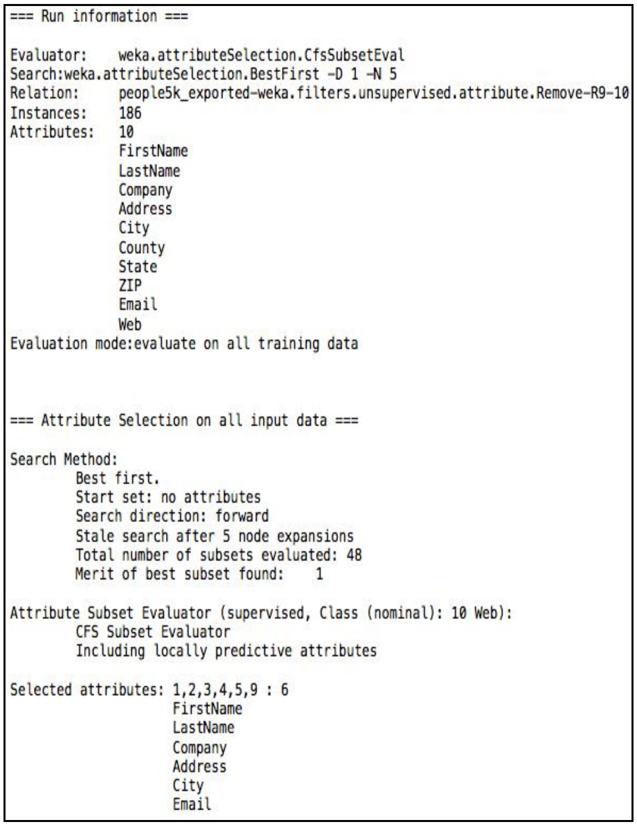
To determine the sensitive data, a priority based approach is used whereby priority is predefined for certain keywords at the time of database setup.

Publicstatic string keywords[] = { "PASS", "ATM", "CARD", "NO","CREDIT","ADDR","ID","DEBIT"};
Publicstatic int priority[] ={ 10,9,8,7,10,6,6,9,10};

Our search for the keywords in the column names in the database and a specific score is determined. If the score is above a threshold, it is deemed sensitive.

The data sets are run through Weka under the Attribute Selection operation. The first is the results from the Pure Data & the second is from the perturbed data. The first shows that the selected Attributes are 5 in number and they are: ("First Name"," Last Name", "Company"," Address"," E-Mail").

The results are pretty much similar; hence the perturbation doesn't significantly affect the data mining results.



```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     people5k_exported-weka.filters.unsupervised.attribute.Remove-R9-10
Instances:    186
Attributes:   10
              FirstName
              LastName
              Company
              Address
              City
              County
              State
              ZIP
              Email
              Web
Evaluation mode:evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 48
        Merit of best subset found:    1

Attribute Subset Evaluator (supervised, Class (nominal): 10 Web):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3,4,5,9 : 6
                     FirstName
                     LastName
                     Company
                     Address
                     City
                     Email
```

Figure 9: Evaluation of Perturbed data

To determine the perturbed data, a best first search method is used to perform the perturbation technique. We use sample data sets of 500, 5000 and 500000 records respectively. The table contains the following attributes/columns like first name, last name, company, address, city, country, state, zip and E-mail.

The data sets are run through Weka under the Attribute Selection operation. The first is the results from the Pure Data & the second is from the perturbed data. The second one shows that the selected Attributes are 6 in number and they are: ("City", "E-Mail"," Company"," Address"," First Name ", " Last Name").

6. CONCLUSION

Privacy is becoming an increasingly important issue in many data mining applications. This has activated the expansion of many privacy-preserving data mining methods. Protecting sensitive raw data in the large database and the knowledge extraction is an important research problem in the field of privacy preserving data mining. In this project, we have protected the sensitive numerical data item in the form of modifying the original data item using the perturbation technique.

The Data Mining results prove that the differences in results between the pure and perturbed data is not significant. Hence this method is successful at preserving privacy during Data Mining. But, there seems to be some loopholes – like for example the email addresses aren't perturbed. This may lead to spam messages and other disturbances, hence the algorithm to identify the sensitive data must be optimized.

Hence, finally we make sure that the data's privacy can be assured and is safely released to any firm or agency for analysis.

REFERENCES

1. Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita. **A review on privacy preserving data mining: techniques and research challenges,** in: *IJCSIT*, Vol. 5, pp. 2310-2315, 2014.
2. M.Z. Islam, L. Brankovic, **Detective: a decision tree based categorical value clustering and perturbation technique in privacy preserving data mining**, *IEEE, INDIN*, 2005.
3. Li Liu, Murat, Bhavani, **The applicability of the perturbation model-based privacy preserving data mining for real world data,** *IEEE, ICDMW*, 2006.
4. S. Vijayarani, A. Tamilarasi, N. Murugesh. **A new technique for protecting sensitive data and evaluating clustering performance**, *IJITCS*, vol. 1, No. 2, April 2011.
5. S. Guha, r. Rastogi, k. Shim **Cure: an efficient clustering algorithm for large databases,** *SIGMOD***,** April 1998.
6. M. Ester, H.P. Kriegel, J. Sander, X. Xu. **A density-based algorithm for discovering clusters in large spatial databases, in:** *KDD*, 1996.
7. Lindell, Yand Pinkas, B., **Privacy Preserving Data Mining**, M. Bellare (Ed.): **Proceedings of the Advances in Cryptology -** *CRYPTO 2000, LNCS 1880, 2000*.
8. R. Kalaivani, S. Chidambaram, **Additive Gaussian noise based data perturbation in multi-level trust privacy preserving data mining,** *IJDKP*, vol. 4, no. 3, May 2014.
9. A. Patel, Samir, **A study of data perturbation techniques for privacy preserving data mining,** *IJSHRE*, vol. 2, issue 2, February 2014.
10. W. Johnson, J. Lipshitz, **Mapping into Hilbert space,** *CM*, May 1984.
11. Agrawal, R. and Srikant, R**., Privacy-preserving Data Mining**, *in Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, Dallas, TX*, May 14-19, 2000.
12. D. Agrawal, and C.C. Aggarwal, **“On the design and quantification of privacy preserving data mining algorithms,”** *Proceedings of the ACM SIGACT–SIGMOD-SIGART Symposium on Principles of Database Systems, ACM New York, NY, USA*, pp. 247-255, 2001.
13. C.K. Liew, U. Choi, C.J. Liew, **A data distortion by probability distribution**, *ACM TODS*, 1985.
14. S.L. WARNER, **A Survey technique for eliminating evasive answer bias,** *JASA*, 1965.
15. Hillo Kargupta and Souptik Dutta, **Random Data Perturbation Techniques and Privacy Preserving Data Mining,** *IEEE,* 2003.