Project Report
on

Privacy Preserving Data Mining

Submitted by

B.BABU(R092749)

A report submitted in partial fulfillment of requirements for the degree

Of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Under the Supervision of

Mr. GAURAV SHRIVASTAV

Lecturer, Department of CSE

RGUKT-RKVALLEY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Rajiv **G**andhi **U**niversity of **K**nowledge **T**echnologies
Idupulapaya, YSR (Dist.), Andhra Pradesh.

BONAFIDE CERTIFICATE

This is to certify that the project report titled **"Privacy Preserving Data Mining"** submitted by **Mr. B BABU** bearing registration numbers **R092749** respectively is a record of bonafide project work carried out under my supervision during the academic year 2014-15.

Signature	**Signature**
Abdul Aziz Md	Mr. Gaurav Shrivastav
(Head Of the Department)	(Lecturer in CSE)
Department of CSE	Department of CSE
RGUKT, RKVALLEY	RGUKT, RKVALLEY

ACKNOWLEDGEMENT

Project is the product out of experience that goes a long way in shaping up a person's caliber. The experience and success one attains is not by oneself, but with a group of kind heart behind.

First and foremost, we express my heartful gratitude to Mr. Gaurav Shrivastav, Lecturer, Dept. of CSE, RGUKT, RK Valley, whose support and encouragement has helped me a lot throughout the progress of my course in research which lead me to complete the project.

We extend my sincere thanks to Mr. Abdul Aziz md, coordinator, Dept. of CSE, RGUKT, RK Valley for providing me with adequate facilities and congenial environment for accomplishing the project.

It is my duty to thank our Vice Chancellor, Prof. Satya Narayana sir, for motivating students towards the fulfillment of project.

Though this is at the end, I heartily express my earnest gratitude to the higher officials of RGUKT and my dear friends, without their lofty inspiration and moral support, this project might not have come as the light of the day.

Yours Sincerely

B. BABU (R092749)

DECLARATION

I certify that

a. The work contained in this report is original and has been done by me under the guidance of my supervisor(s).
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in preparing the report.
d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Signature of the Student

Table of Contents

1. Introduction

1.1 Motivation :

Recent enhancements in information science have made possible the collection and analysis of billions of transactions containing personal sensitive information. These data include criminal records, shopping habits, medical histories, credit records and others. This development in the storage and analysis of data has led individuals and organizations to face the challenge of turning such data into useful information and knowledge.

Data mining is a correct approach to reach this challenging necissities. The arena of data mining, also called Knowledge Discovery in Databases (KDD), has got a special attention since the 1990s. This new research domain has came out as a means of getting(extracting) hidden models or previously unknown implicit patterns from large databases. The fascination with the promise of analysis of large volumes of data has led to an increasing number of successful applications of data mining in recent years. Undoubtedly, these applications are very useful in many areas such as marketing, business, medical analysis, and other applications in which pattern discovery is paramount for strategic decision making.

Despite its uses in various areas, the use of data mining approaches can also result in new threats to privacy and information security. These issues, such as those involved in privacy-preserving data mining (PPDM), cannot controlled by simply restricting data collection or even by restricting the secondary use of this huge amounts of data. Moreover, there is no exact solution that solves privacy preservation in data mining. An approximate solution could be sufficient, depending on the application since the appropriate level of privacy can be interpreted in different contexts.

Preserving privacy when data are published for mining is a considerably challenging problem. The general methods in database security, such as access control and authentication that have been adopted to successfully manage the access to data present some limitations in the context of data mining. While access control and authentication protections can guard against direct disclosures, they do not address disclosures based on inferences that can be drawn from released data. Preventing this type of inference detection is beyond the reach of the existing methods.

1.2 Privacy Preserving : Problem Definition

In this report, we address the problem of converting a database into a new one that secures sensitive information while preserving the general patterns and trends from the original data. The sensitive information is not limited to personal data, but may reflect customers' purchasing behaviour, financial, medical, and insurance information and sensitive patterns.

The problem of protecting sensitive knowledge in transactional databases, draws the following assumptions:

1. Data owners have to know in advance some knowledge (rules) that they want to protect. Such rules are fundamental in decision making, so they must not be discovered.
2. The individual data values (e.g. a specific item) are not restricted. Rather, some aggregates and relationships must be protected. This approach works in the opposite

way to the idea behind statistical databases which prevents against discovering individual tuples.

1.3 Project : Statement

This research demonstrates empirically and theoretically the practicality and feasibility of achieving PPDM. Three major issues are addressed to support the central thesis(report) statement of this research, as follows:

1. It is possible to transform a database by protecting the attribute values of objects subjected to clustering and get valid clustering results, i.e., the clusters generated in the transformed database are very similar to those mined from the original database.
2. It is possible to protect sensitive knowledge discovered from databases without losing the benefit of mining the transformed database.
3. It is possible to quantify the disclosure of sensitive knowledge discovered from a transformed database.

1.4 Research Methodlogy :

First we have studied all most all basic concepts in Data mining course. We had been analysing almost 50 IEEE papers. And verified some thesises of Ph.D students. Finally we come with new(weighted attributes) clustering based anonymization technique.

1.5 Organization of Report :

The rest of this dissertation is organized as follows :

Next part contains background knowledge for privacy preserving data mining. Here we had explained mainly why we need privacy in data mining and some techniques to preserve privacy(its advantages and disadvantages). Third part contains proposed work. This pard explains our algorithm and weight assignments. Next part contains Conclusion and finally references.

2. Basic Concepts

2.1 About Data

Data is a collection of records(tuples) from an organization database. The number of tuples ranges of upto some millions.

Mainly there are three types of attributes.

1. **Personal attributes :**

   Attributes which identifies a person uniquely are called "**Personal Attributes**". Like Name, Address(complete), PAN card number, Bank Account Number. These attributes simply leases information about a individual, so we have to remove these attributes completely while publishing data.

2. **Quasi identifiers :**

   A set of attributes which identifies a person partially, means gives set of persons details instead of single person. Sometimes in worst case it gives exact person details also. So we need to anonymize quasi identifiers such that user can't access the personal information.

3. **Sensitive attributes :**

   Attributes which are confidentail. Our main goal is to protecting these sensitive attributes.

2.2 Frameworks for Protecting Privacy

There are broadly two frameworks for protecting privacy in statistical databases. In the interactive framework, the user (researcher) queries the database through a privacy mechanism, which may deny the query or alter the answer to the query in order to ensure privacy. In the non-interactive framework, the original database is first sanitized so as to preserve privacy and then the modified version is released. Interactive framework mainly comprises of two methods: query auditing and output perturbation. In query auditing, a query is denied if the response could reveal sensitive information and answered exactly otherwise. In output perturbation, the privacy mechanism computes the exact answer to the query and then outputs a perturbed version (say, by adding noise) as the response to the query. The methods under non-interactive framework generally involve perturbation of the data (input perturbation), anonymization (k-Anonymity), computing summaries, or a combination of these. In input perturbation, the original database is perturbed (say, by adding noise) into a transformed database, which is then released to the users. In k-Anonymity, the identifying fields are removed first and then some of the remaining entries are suppressed or generalized so that for each tuple in the modified table, there are at least $k - 1$ other tuples identical to it. Yet another approach, that is somewhat orthogonal to the above classification, is to use techniques from secure multi-party computation. Using these techniques, several parties with private inputs can compute a function of their inputs such that an adversarial party (or a coalition of parties) cannot learn any information that cannot be deduced from the output of the function and the input of the adversary. For example, two hospitals may want to learn the correlation between age and heart disease across their patient databases without revealing any other information.

2.3 Relative Merits of Different Privacy Preserving Techniques

We next emphasize that both the frameworks are relevant in different contexts. The results obtained in the interactive framework are expected to be of better quality since only queries of interest to the user are answered, whereas in the non-interactive framework, the user has access to the entire sanitized database and hence can compute answer to any query. For example, under many circumstances, the results obtained using output perturbation are of provably better quality than is possible for non-interactive solutions. On the other hand, non-interactive methods are preferable whenever the underlying data mining task is inherently ad hoc and the researchers have to examine the data in order to discover data aggregation queries of interest. Moreover the sanitization can be done offline as no interaction with the user is needed. We can also avoid the risk of accidental or intentional disclosure of the sensitive data by deleting the original data or locking it in a secure vault. Further, for all interactive methods, collusion and denial of service are problems of concern. There is an implicit assumption that all users can collude with each other and hence queries from all users are treated as coming from a single user. Consequently any one user has reduced utility. In particular, a malicious user may pose queries in such a way that many innocuous queries are either denied or answered with excessive noise (as the case may be) in the future.

Similarly each of the above methods has its own advantages and disadvantages and depending on the application some method may be better than others. While query auditing and non-interactive methods maintain consistency (i.e., if the same query is posed again, we get the same answer), output perturbation does not. The query auditing method is useful in settings where exact answers to queries are necessary. For example, doctors often require exact answers to queries when designing new drugs. Similarly, among the non-interactive methods, k-Anonymity method is desirable when we want to draw inferences with 100% confidence. Secure function evaluation is useful when the aggregate function (such as set intersection) is known a priori and there are efficient protocols for computing this function. However, privacy is preserved only to the extent that the output of the function itself does not reveal any private information. The difficult part is to determine the functions that are privacy-preserving in the context of statistical databases.

2.3.1 Privacy Protection in the Interactive Framework :

-Query Auditing

Consider a data set consisting of private information about individuals. The online query auditing problem is: given a sequence of queries that have already been posed about the data, their corresponding answers and given a new query, deny the answer if privacy can be breached or give the true answer otherwise. Attacker can compromise the privacy of a large fraction of the individuals in the data because of small fundamental problems of these algorithms. To overcome this problem, there is a new model called simulatable auditing where query denials provably do not leak information.

-Output Perturbation

Another dimension along which the privacy techniques can be classified is the amount of trust required on the database administrator. The positive results in the privacy literature fall into three broad categories: non-interactive with trusted server, non-interactive with untrusted server – specifically, via randomized response, in which a data holder alters her data with some probability before sending it to the server – and interactive with trusted server. In particular, the privacy methods for the interactive framework assume that the database administrator is trusted by the individuals whose private information is contained in the database. Inspired by the desire to enable individuals to retain control over their information, There are some distributed implementation of the output perturbation schemes described in different literatures, thereby removing the assumption of a trusted collector of data. Such an approach is desirable even from the perspective of an organization such as census bureau: the organization does not have to protect against insider attacks or worry about the high liability costs associated with a privacy breach.

There are some implementation replaces the trusted server with the assumption that strictly fewer than one third of the participants are faulty. In the above output perturbation schemes, privacy is obtained by perturbing the true answer to a database query by the addition of a small amount of Gaussian or exponentially distributed random noise. Under many circumstances the results obtained are of provably better quality (accuracy and conciseness, i.e., the number of samples needed for correct statistics to be computed) than is possible for randomized response or other non-interactive solutions.

2.3.2 Privacy Protection in the Non-interactive Framework

- K-Anonymity

Next we consider the problem of releasing a table containing personal records, while ensuring individual privacy and maintaining data integrity to the extent possible. When the aggregate queries of interest are not known a priori, techniques such as query auditing, output perturbation, and secure function evaluation do not provide an adequate solution, and we need to release an anonymized view of the database that enables the computation of non-sensitive query aggregates, perhaps with some error or uncertainty. Moreover, techniques under non-interactive framework such as input perturbation, sketches, or clustering may not be suitable if one wants to draw inferences with 100% confidence. Another approach is to suppress some of the data values, while releasing the remaining data values exactly. We note that suppressing just the identifying attributes, such as name and social security number, is not sufficient to protect privacy. This is because we can still join the table with public databases (such as voter list) and identify individuals using non-identifying attributes, such as age, race, gender, and zip code (also called quasi-identifying attributes). In order to protect privacy, we adopt the k-Anonymity model which was proposed by Samarati and Sweeney. Suppose we have a table with each tuple having only quasi-identifying attributes. In the k-Anonymity model, we suppress or generalize some of the entries in the table so as to ensure that for each tuple in the modified table, there are at least $k - 1$ other tuples in the

modified table that are identical to it. Consequently, even with the knowledge of an individual's quasi-identifying attributes, an adversary cannot track down an individual's record further than a set of at least k records. In other words, releasing a table after k-anonymization keeps each individual hidden in a crowd of k − 1 other people. We study the problem of k-Anonymizing a table, with minimum amount of suppression/generalization and provide approximation algorithms for it.

EXAMPLE :

Input data

NONSENSITIVE			SENSITIVE
PID	**STATE**	**AGE**	**DISEASE**
121045	Odisha	27	Brain Cancer
121077	Bihar	28	Malaria
121088	UP	29	Heart Disease
121067	MP	22	Malaria
134222	Odisha	55	Heart Disease
134567	MP	54	Malaria
134889	UP	50	Heart Disease
134778	Bihar	47	Malaria
143367	Odisha	35	Heart Disease
148000	MP	37	Brain Cancer
145690	UP	33	Malaria
148056	Bihar	34	Brain Cancer

Output data :

NONSENSITIVE			SENSITIVE
PID	**STATE**	**AGE**	**DISEASE**
1210**	*	<30	Brain Cancer
1210**	*	<30	Malaria
1210**	*	<30	Heart Disease
1210**	*	<30	Malaria
134***	*	>40	Heart Disease
134***	*	>40	Malaria
134***	*	>40	Heart Disease
134***	*	>40	Malaria
14****	*	3*	Heart Disease
14****	*	3*	Brain Cancer
14****	*	3*	Malaria
14****	*	3*	Brain Cancer

- Anonymity via Clustering

Consider the problem of publishing data for analysis from a table containing personal records, while maintaining individual privacy. We propose a new method for anonymizing data records, where quasi-identifiers of data records are first clustered and then cluster centers are published. To ensure privacy of the data records, we impose the constraint that each cluster must contain no fewer than a pre-specified number of data records. This technique is more general since we have a much larger choice for cluster centers than k-Anonymity. In many cases, it lets us release a lot more information without compromising privacy. By not releasing a small fraction of the database records, we can ensure that the data published for analysis has less distortion and hence is more useful. Our approximation algorithms for new clustering objectives are of independent interest and could be applicable in other clustering scenarios as well.

3. Proposed Work

3.1 Necissity of new Algorithm

We have few disadvantages in k-anonymity like l-diversity, t-closeness. If we want to get rid of these disadvantages we have to apply some contraints on k-anonymity which reduces information. And as a result data mining results looses efficiency. So we come with a new approach which can preserve our data without any weaknesses. K-anonymity, other alternatives methods have approximation algorithms(NP-complete). Our algorithm is polynomial time algorithm. In our algorithm we follow k-mean clustering algorithm with few modifications for anonymizing data. In proposed clustering method, feature weights are manual assigned so that the information distortion can reduced.

Clustering aims at grouping a set of tuples into a group so that objects in each group are similar to each other and are different from tuples in other clusters. In the k-anonymity protected data, if the tuples that will be grouped as an equivalence class are more similar to each other, it reveals the more sensitive information distortion for generalizing the equivalence class. This is the reason why the k-anonymity model can be addressed from the viewpoint of clustering. If number of records are more(millions or billions) then k-anonymity model is inefficient.

3.2 Algorithm

Input : A table T contains M records in which each record has N quasi-identifier features and the value of k in the k-anonymity model.(after data selection, data transformation)

Output : Anonymized table

Algorithm :

1. Assign waits for every attribute(sum of weights should be one).
2. Calculate number of initial clusters(#clusters(c) = M / k).
3. Assign initial random centroids for every cluster.
4. For every tuple
   - calculate distance(d_i) between every cluster centroid.
   - assign that tuple to minimum distance cluster.
   - update centroid(simple averaging).
5. For every cluster
   - if cluster contains morethan k tuples.
     - publish that centroid, number of records and sensitive information.
   - otherwise ignore that cluster

Complexity : $M \times c$

3.3 Example

3.3.1 Input :

Dataset : adult dataset

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

```
39,State-gov,77516,Bachelors,13,Never-married,Adm-clerical,Not-in-family,White,Male,2174,0,40,United-States,<=50K
50,Self-emp-not-inc,83311,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,13,United-States,<=50K
38,Private,215646,HS-grad,9,Divorced,Handlers-cleaners,Not-in-family,White,Male,0,0,40,United-States,<=50K
53,Private,234721,11th,7,Married-civ-spouse,Handlers-cleaners,Husband,Black,Male,0,0,40,United-States,<=50K
28,Private,338409,Bachelors,13,Married-civ-spouse,Prof-specialty,Wife,Black,Female,0,0,40,Cuba,<=50K
37,Private,284582,Masters,14,Married-civ-spouse,Exec-managerial,Wife,White,Female,0,0,40,United-States,<=50K
49,Private,160187,9th,5,Married-spouse-absent,Other-service,Not-in-family,Black,Female,0,0,16,Jamaica,<=50K
52,Self-emp-not-inc,209642,HS-grad,9,Married-civ-spouse,Exec-managerial,Husband,White,Male,0,0,45,United-States,>50K
31,Private,45781,Masters,14,Never-married,Prof-specialty,Not-in-family,White,Female,14084,0,50,United-States,>50K
42,Private,159449,Bachelors,13,Married-civ-spouse,Exec-managerial,Husband,White,Male,5178,0,40,United-States,>50K
37,Private,280464,Some-college,10,Married-civ-spouse,Exec-managerial,Husband,Black,Male,0,0,80,United-States,>50K
30,State-gov,141297,Bachelors,13,Married-civ-spouse,Prof-specialty,Husband,Asian-Pac-Islander,Male,0,0,40,India,>50K
23,Private,122272,Bachelors,13,Never-married,Adm-clerical,Own-child,White,Female,0,0,30,United-States,<=50K
32,Private,205019,Assoc-acdm,12,Never-married,Sales,Not-in-family,Black,Male,0,0,50,United-States,<=50K
40,Private,121772,Assoc-voc,11,Married-civ-spouse,Craft-repair,Husband,Asian-Pac-Islander,Male,0,0,40,?,>50K
34,Private,245487,7th-8th,4,Married-civ-spouse,Transport-moving,Husband,Amer-Indian-Eskimo,Male,0,0,45,Mexico,<=50K
25,Self-emp-not-inc,176756,HS-grad,9,Never-married,Farming-fishing,Own-child,White,Male,0,0,35,United-States,<=50K
32,Private,186824,HS-grad,9,Never-married,Machine-op-inspct,Unmarried,White,Male,0,0,40,United-States,<=50K
38,Private,28887,11th,7,Married-civ-spouse,Sales,Husband,White,Male,0,0,50,United-States,<=50K
43,Self-emp-not-inc,292175,Masters,14,Divorced,Exec-managerial,Unmarried,White,Female,0,0,45,United-States,>50K
40,Private,193524,Doctorate,16,Married-civ-spouse,Prof-specialty,Husband,White,Male,0,0,60,United-States,>50K
54,Private,302146,HS-grad,9,Separated,Other-service,Unmarried,Black,Female,0,0,20,United-States,<=50K
35,Federal-gov,76845,9th,5,Married-civ-spouse,Farming-fishing,Husband,Black,Male,0,0,40,United-States,<=50K
43,Private,117037,11th,7,Married-civ-spouse,Transport-moving,Husband,White,Male,0,2042,40,United-States,<=50K
59,Private,109015,HS-grad,9,Divorced,Tech-support,Unmarried,White,Female,0,0,40,United-States,<=50K
56,Local-gov,216851,Bachelors,13,Married-civ-spouse,Tech-support,Husband,White,Male,0,0,40,United-States,>50K
19,Private,168294,HS-grad,9,Never-married,Craft-repair,Own-child,White,Male,0,0,40,United-States,<=50K
54,?,180211,Some-college,10,Married-civ-spouse,?,Husband,Asian-Pac-Islander,Male,0,0,60,South,>50K
39,Private,367260,HS-grad,9,Divorced,Exec-managerial,Not-in-family,White,Male,0,0,80,United-States,<=50K
49,Private,193366,HS-grad,9,Married-civ-spouse,Craft-repair,Husband,White,Male,0,0,40,United-States,<=50K
23,Local-gov,190709,Assoc-acdm,12,Never-married,Protective-serv,Not-in-family,White,Male,0,0,52,United-States,<=50K
20,Private,266015,Some-college,10,Never-married,Sales,Own-child,Black,Male,0,0,44,United-States,<=50K
45,Private,386940,Bachelors,13,Divorced,Exec-managerial,Own-child,White,Male,0,1408,40,United-States,<=50K
30,Federal-gov,59951,Some-college,10,Married-civ-spouse,Adm-clerical,Own-child,White,Male,0,0,40,United-States,<=50K
22,State-gov,311512,Some-college,10,Married-civ-spouse,Other-service,Husband,Black,Male,0,0,15,United-States,<=50K
48,Private,242406,11th,7,Never-married,Machine-op-inspct,Unmarried,White,Male,0,0,40,Puerto-Rico,<=50K
```

3.3.2 Output

Anonymity level : 3
Relavent Attributes : 9/15
Output data : contains 90% of data

```
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,Yes
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,Yes
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
39,Local-gov,12,Divorced,Adm-clerical,Not-in-family,Male,34,No
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
42,Private,13,Widowed,Sales,Not-in-family,Male,60,Yes
```

4. Conclusion

Publishing data about individuals without revealing sensitive information is an important problem. The notion of privacy called k-Anonymity has attracted a lot of research attention recently. In a k-anonymized database, values of quasi-identifying attributes are suppressed or generalized so that for each record there are at least k-1 records in the modified table that have exactly the same values for the quasi-identifiers. However, the performance of the best known approximation algorithms for k-Anonymity depends linearly on the anonymity parameter k. In this project, we introduced clustering as a technique to anonymize quasi-identifiers before publishing them. We have implemented weighted distance clustering approach for anonymization. This method is free from all most all attacks, its output does not depends on input data size.

5. References

1. Pierangela Samarati, Latanya Sweeney : "*k-anonymity and its enforcement through generalization and suppression*".
2. Qian Wang, Zhiwei Xu and Shengzhi Qu : "*An enhanced K-anonumity model against homogeneity attack*", 2011.
3. Dan Zhu, Xiao-Bai Li, Shuning Wu : "*Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining*". Decision Support Systems. December 2009, Volume 48, Pages 133-140.
4. Heiko Paulheim : "Exploiting linked open data as background knowledge in data mining".
5. Debasis Mohapatra1, Dr. Manas Ranjan Patra : "*Analysis of k-Anonymity for Homogeneity Attack*", International Journal of Advances in Computer Science and Technology, 2014
6. Tiancheng Li, Ninghui Li : "*Mining background knowledge for data anonymization*".
7. Ashwin Machanacajjhala, Johnnes Gehrke : *"Daniel Kifer privacy beyond k-anonymity"*.
8. Md. Zahidul Islam and Ljiljana Brankovic *: "A Framework for Privacy Preserving Classification in Data Mining", 2004, Australian Computer Society.*
9. Arik Friedman and Assaf Schuster *: "Data Mining with Differential Privacy"*.
10. Jim Dowd, Shouhuai Xu, and Weining Zhang *: "Privacy-Preserving Decision Tree Mining Based on Random Substitutions"*.
11. Shucheng Yu, Cong Wang, Kui Ren, *"Attribute Based Data Sharing with Attribute Revocation"*,ASIACCS'10 April 13–16, 2010, Beijing, China.
12. Batya Kenig, Tamir Tassa, *"A Practical Approximation Algorithm for Optimal k-Anonymity"*.
13. Alexandre Evfimievski, *"Randomization in Privacy Preserving Data Mining"*
14. Stanley Robson de Medeiros Oliveira, "*Data Transformation For Privacy-Preserving Data Mining*", PhD thesis, University of Alberta, 2005.
15. Nan Zhang, "*Privacy preserving data mining*", Phd Thesis, Texas A&M University
16. archive.ics.uci.edu/ml/datasets/Adult

Appendix A : System Structure / Developer Manual

This is whole system is developed by using simple Core Java operations. Most of this project code contains about reading and writing data from/to file. This code contains mainly three packages.

1. **Preprocessing**

   This package performs first few steps of data mining. It removes unwanted data(attributes) from input data file, transforms whole data into numerical data.

2. **Anonymity**

   This package performs clustering alogrithm with weights. And finally it transforms numerical data into corresponding data.

3. **Utility**

   Contains operations related to utilities like reading/writing file, etc

Apart from these Java components, we have one directory called **data** in our system. In that we have many simple text files

i. **input data file (ex : adultset.data)** : Contains input data

ii. **output.txt** : Stores output data, for which we are built this system.

iii. **missingValues.txt** : User has to enter missing values of each relevvant attributes.

iv. **transformedData.txt :** This is temporary file to store transformed data(between preprocessing and clustering) this file is very useful when our input data contains more than 10^6 tupules because we can't store those tuples in arrays.

v. **metaData.config** : This configuration file is very important for whole system because this contains each and every aspect of system. Developer has to take dataset specific information from this config file.

Appendix B : System Structure / User Manual

If user want to use this system first user has to modify metaData.config as per the dataset.

Instructions :

- First line contains exact path of input file.
- Second line contains separator of attribute values in input file.
- Third line contains missing value symbol(signal).
- Next line contains attribute description(ex : 0112012). This description contains only 0,1,2 digits. '0' means numerical attributes . '1' means categorical attributes. '2' means irrevalent attribute.
- Next few lines contains information about attribute ranges, and equivalent numerical values for categorical values. Like 'a:2' means 'a' has equivalent numerical value 2.(see example). For numerical attributes it contains minimum value, maximum value and interval difference.
- Each attribute description takes one line descritption.
- After that anonymity level
- Total tuples
- Weights for each attribute. Order of weights matters a lot.

Example of description file :

./data/adultset.data

,

?

012201112122021

17 90 1

8 Federal-gov:1 State-gov:2 Local-gov:3 Private:4 Self-emp-not-inc:5 Self-emp-inc:6 Without-pay:7 Never-worked:8

1 16 1

7 Never-married:1 Divorced:2 Separated:3 Widowed:4 Married-civ-spouse:5 Married-spouse-absent:6 Married-AF-spouse:7

14 Prof-specialty:1 Tech-support:2 Armed-Forces:3 Sales:4 Exec-managerial:5 Craft-repair:6 Machine-op-inspct:7 Handlers-cleaners:8 Adm-clerical:9 Transport-moving:10 Farming-fishing:11 Protective-serv:12 Priv-house-serv:13 Other-service:14

6 Unmarried:1 Not-in-family:2 Husband:3 Wife:4 Own-child:5 Other-relative:6

2 Male:1 Female:2

1 100 1

2 >50K:1 <=50K:0

3

32561

0.11 0.15 0.2 0.05 0.25 0.02 0.02 0.2