PRESERVING THE SENSITIVE INFORMATION USING HEURISTIC BASED APPROACH

ANIKET PATEL, NISHA KHURANA

Abstract: Privacy preserving is becoming an important issue in the field of data mining. Privacy preserving in data mining is gaining popularity day by day as it allows sharing of privacy-sensitive data which is used for analysis purpose. People find it uncomfortable to share their sensitive information hence they either provide wrong data or simply refuse to share. As a result due to incorrect data collection affect the success of data mining as it relies on significant amount of accurate data so that meaningful and correct results may be produced. In recent years, the wide availability of personal data has made the problem of privacy preserving data mining an important one. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper reiterates several privacy preserving data mining technologies and then analyzes the merits and shortcomings of these technologies. Using such approaches the data accuracy and preservation can be achieved. An effective approach is heuristic based approach that has aim to achieve the privacy of static data with minimum information loss.

Keywords*:* Data Mining, Privacy preserving data mining, heuristic based approach, ARX Tool

Introduction: Recent years have seen tremendous advancement in hardware technology leading to the increased capability of saving and recording personal data about consumers and individuals. Many business organizations collect data for analyzing business policy of competitors, consumer behavior and improving their own business strategies. Large amount of such data can be used for different data mining purposes such as knowledge discovery, decision-making, statistical analysis etc. There are different perspectives of privacy to different people. Some may consider entire personal information as private while some may think certain attribute value should not be available directly or indirectly to the personal domain [4].

Privacy preserving data mining (PPDM) algorithms attempt to reduce the injuries to privacy caused by malicious parties during the rule mining process. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. All privacy-preserving transformations cause information loss, which must be minimized in order to maintain the ability to extract meaningful information from the published data. Despite such recognition and efforts, there
remain many challenges to foster data mining task while protecting individual privacy.

Privacy preserving Data Mining: Generally, when people talk of privacy, they mean original information about them should not be available to others. It is this intrusion, or use of personal data in a way that negatively impacts someone's life and is the cause of concern. As long as data is not misused, most people do not feel their privacy has been violated. The problem is that once information is released; it may be difficult to prevent misuse. Utilizing this distinction, ensuring that a data mining would not enable misuse of personal information, opens opportunities to implement privacy in a number of ways. To do this, technical
and social solutions that ensure data will not be released. The same basic concerns also apply to collection of data.

Given a collection of data, it is possible to learn things that are not revealed by any individual data item. An individual may not care about someone knowing their birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. This type of privacy problem arises with large, multi-individual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such corporate information is generally the goal of data mining, but some results may still lead to concerns (often termed a secrecy, rather than privacy issue.) The difference between such corporate privacy issues and individual privacy is not that significant, if we view disclosure of knowledge about an entity (information about an individual) as a potential individual privacy violation.

Privacy Preserving Data Mining (PPDM) Techniques: Various techniques have been proposed varying from entire data set modification to selective data set modification. K-anonymity offers selective data set modification to reduce granularity of the data using techniques such as generalization and suppression. Privacy preserving data mining techniques can be classified in three groups. Namely Reconstruction based technique, Heuristic based technique, and Cryptography based technique.

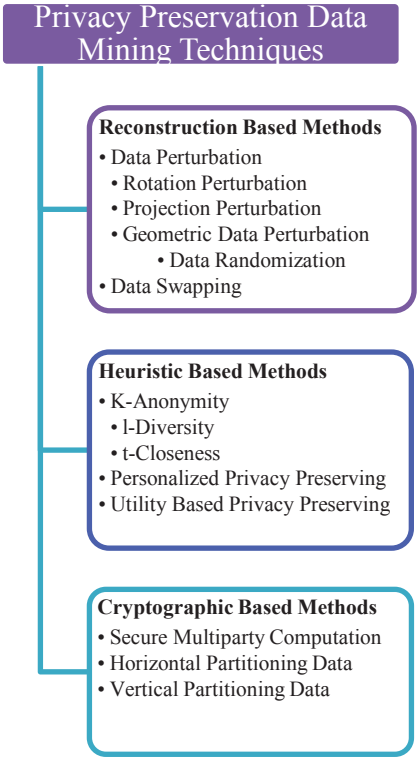

Figure 1: Classification of Privacy Preserving Data Mining [15]

Related Work:

Reconstruction based technique: Reconstruction-based techniques perturb the original data to achieve privacy preserving. The perturbed data would meet the two conditions. First, an attacker cannot discover the real original data from the issuance of the distortion data. Second, the distorted data is still to maintain some statistical properties of the original data, namely some of the information derived from the distorted data are equivalent to data obtained from the original information [6]. So for each data mining technique separate algorithms need to be developed. Data perturbation is that the value of each property is transformed into other value of the property domain by given probability .

Data perturbation: Basically change the original data into another.

- **Rotation perturbation:** Rotate the data according to a given angle θ with the origin as the center[1][2].
- **Projection perturbation:** Projecting a set of data points from the original multidimensional space to another randomly chosen space[1].
- **Geometric perturbation:** Transform the data into sequence of random geometric transformations[1].

Data randomization: Randomly choose the data from the record and place it at the place of another data.[3]

Data swapping: Swaps the data into two halves [2].

Heuristic based approach: Various techniques have been developed to sanitize or modify selective data for data mining techniques like association rule mining, classification and clustering. Selective data sanitization or modification based mining problem is NP-hard and for this reason, heuristics can be used to address the complexity issues. The concept of protecting respondent identity through micro data release using k-anonymity was first proposed by P. Samarati in [9], and subsequently many techniques have been proposed based on it, such as l-diversity [15], t-cıoseness [8], Incognito [7], and so on. K-anonymity protects against identity disclosure; it does not provide sufficient protection against attribute disclosure.

We know that the database contains different types of attributes. Explicit Identifiers: are attributes which can identify the person uniquely, e.g identity number,

PAN card number etc. [10] Quasi-Identifier (QI): The attributes which cannot alone identify the person uniquely but by collecting quasi-identifiers any one can easily recognize any person, example Zip-code, Birth-date etc. Sensitive Attributes (SA): The information which everyone tries to hide from adversaries example salary, disease etc[11][12]. Non-Sensitive Attributes: The attributes which does not relate to any other categories and have no importance when disclosed to anyone. Each group that shares the same values on every QI is called Equivalence Class (EC). While releasing the sensitive information it is required to preserve them from disclosure. There are mainly two types of Information Disclosure: Identity Disclosure and Attribute Disclosure. [11][12][13][14].

k-Anonymity: The database is said to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. K-Anonymity thus prevents definite database linkages and guarantees that the data released is accurate. One of the emerging concept in microdata protection is *k-anonymity*, which has been recently proposed as a property that captures the protection of a microdata table with respect to possible re-identification of the respondents to which the data refer. *K*-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than *k* respondents.

l-Diversity: In this concept, say you have a group of k different records that all share a particular quasi-identifier. That's good, in that an attacker cannot identify the individual based on the quasi-identifier. But what if the value they're interested in, (e.g. the individual's medical diagnosis) is the same for every value in the group. The distribution of target values within a group is referred to as "*l*-diversity". [16]

t-Closeness: *t*-closeness that formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*). This effectively limits the amount of individual-specific information an observer can learn. Intuitively, privacy is measured by the information gain of an observer.

Personalized privacy preservation: minimize the generalization for satisfying everyone's requirements, so discard the maximum amount of information from the microdata (raw data is called micro data). [18]

Utility based privacy preservation: To improve the query answering accuracy on anonymized table. [19]

Cryptography based technique: In a distributed environment the primary issue to achieve privacy preserving is the security of communications, and encryption technology just to meet this demand. Therefore, privacy preserving based on data encryption technology commonly applies to distributed applications. Lindell & Pinkas [17] first proposed Secure Multi-Party Computation protocol for data mining classification techniques. Cryptography based techniques offer a well-defined model for privacy, which includes methodologies for proving and quantifying it. Cryptography-based techniques have more time complexity compare to other method for data updating. Cryptography techniques are used to preserve privacy.

Distributed data mining provides different algorithms to perform computation in distributed manner without pooling the whole data into one place.

The secure multiparty computation is one of the distributed computing examples which is using worldwide for data distributed across the network. Firstly Yao[9] introduced the secure multiparty computation technique.

Table 1: Comparison table of techniques

Technique	**Advantages**	**Disadvantages**
Reconstruction Based Technique	Data is transformed to achieve greater security. Different attributes are preserved independently.	Reduce the granularity loss of effectiveness of data.
Cryptography Based Technique	Encryption provide security to data.	More time complexity Security and attacks. Difficult to scale multiple parties are involved.
Heuristic Based Technique	Handle data in group based manner	Handling sensitive Data. Linking Attacks.

I. HEURISTIC BASED APPROACH FOR PRIVACY PRESERVING STATIC DATA MINING.

Framework: Here figure 2 describes the framework of heuristic based approach for privacy preserving static data mining. The static dataset D is used to create Semantic Hierarchy tree of different attributes like sensitive attributes, quasi attributes etc... Then heuristic based algorithm K-anonymity and l-diversity is applied on the attributes using semantic hierarchy tree. Data classification and clustering algorithm is applied and then the result is produced D'' and then compare to the original dataset D'.

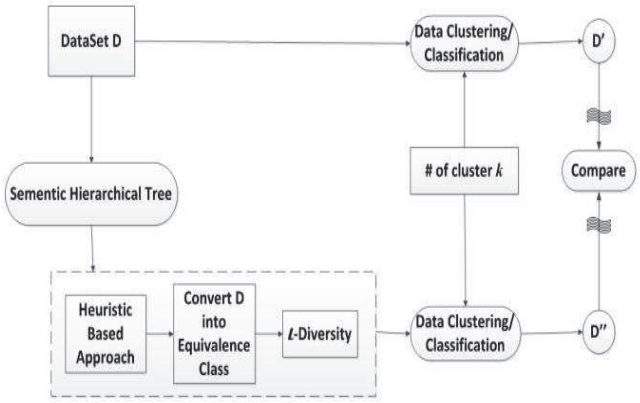

Figure 2: Framework for heuristic based approach for privacy preserving static data mining

Algorithm

Procedure: Heuristic based approach for Privacy Preserving Static Data D(x):

Input: The original Data Stream D; the number of quasi- identifier attributes q; equivalence class size k; similarity measure ε; Sensitive attributes a, Number of sensitive attributes n;

Output: Satisfy k-anonymous anonymity table D'.

Step 1: Convert the Quasi-identifiers into semantic Hierarchical tree for classification. Let t= {v1, v2,..., vk } be a tuple with k quasi-identifier values and t'={ v1', v2',..., vk'} be a generalized tuple of t.Step 2: Form initial equivalence class, S={E1, E2, E3,..., En}Step 3: While T exists an equivalence class of size < k do

(i) Select an equivalence class Ei of size <k;(ii)Calculate the pair wise distance between Ei and remaining equivalence classes in T, find the equivalence class Ej with the smallest distance to ECi, and multiple sensitive attribute values are not ε-similar in ECi∪ EC j.Step 4: For multiple sensitive attributes n in S the smallest distance between EC1n and EC2n is consider using l- diversity.Step 5: Then for ECi ∪ EC j (i) if (| ECi | + | EC j|) < 2then ECi = ECi ∪ EC j ; T = T – EC j ;

(ii) if (| ECi | + | EC j|) ≥ 2k then

Select k-|ECi| tuples from ECj and merge them with the smallest distance Ei to form a lager equivalence class with sensitive attribute values are not mutually ε-similar.

Step 6: Apply k-Mean clustering algorithm with different values of k on original dataset D having sensitive attribute a.

Step 7: Apply k-Mean clustering algorithm with different values of k on Anonymized dataset D' having sensitive attribute a

Step 8: Create cluster membership matrix of results from step 6 and step7 and analyze.

Experiments were performed to measure the accuracy of data while protecting the sensitive attributes at the same time. Here we use ARX Tool to calculate or measure the accuracy of data. This tool is one of the powerful anonymization tool. In this tool, analysis risk is a perspective where various analysis risks are measured on the data.

It includes re-identification risks for the user that specify what is the measure of risk in disclosure of data.

Table 2: Dataset attributes description

Dataset= 30162	
Attributes	Types
Sex	Quasi-attribute
Age	Quasi-attribute
Race	Sensitive attribute
Marital-status	Sensitive attribute
Education	Quasi-attribute

Native-country	Sensitive attribute
Work-class	Quasi-attribute
Occupation	Quasi-attribute
Salary-class	Sensitive attribute

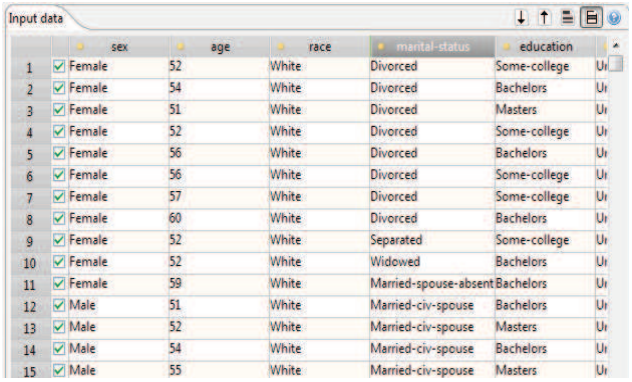

Figure 3: Original Data D

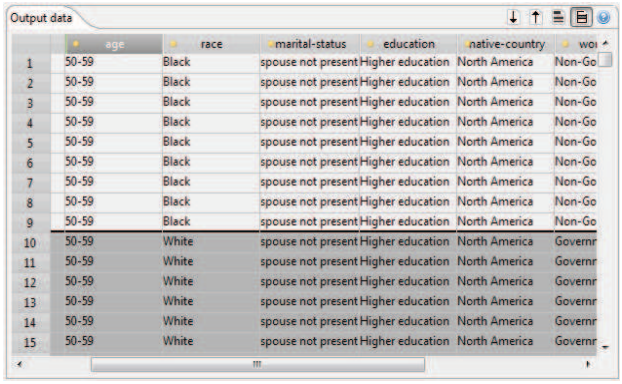

Figure 4: 3-Anonymized Data D'.

Re-identification risks: Re-identification risks may be analyzed based on sample characteristics or on the concept of uniqueness. Uniqueness can either be determined based on the sample itself or it may be estimated with super-population models.

Here risk estimates provide three different attackers' models: [20]

The prosecutor scenario,

The journalist scenario and

The marketer scenario

In **Prosecutor Model,** the attacker already knows that data about the targeted individual is contained in the data set.

In **Journalist Model**, the background knowledge is not assumed.

In **marketer Model**, the attacker is not interested in re-identifying a specific individual but that she aims at attacking a larger number of individuals.

The Threshold value is provided for the analyzing highest risk of any record, the records that have a risk higher than this threshold and for the average proportion of records that can successfully be re-identified.

They are complemented by numbers on population uniqueness from a selected statistical model:

-Lowest prosecutor re-identification risk.

-Individuals affected by lowest risk.

-Highest prosecutor re-identification risk.

-Individuals affected by highest risk.

-Average prosecutor re-identification risk

Distribution of equivalence class sizes: In this view, the distribution of sizes of equivalence classes (or cells) can be analyzed. The distribution is displayed for both input and output data, either as a histogram or as a table.

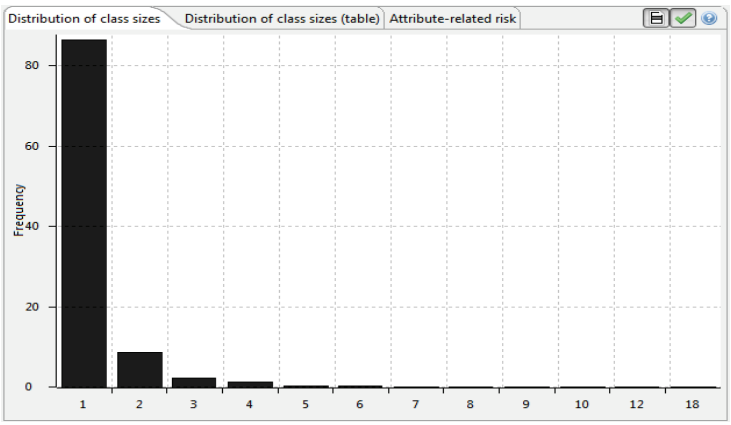

Figure 5: Distribution of class size of original data D.

In anonymization technique the data is distributed in equivalence class. This diagram shows the distribution of equivalence class of original data.

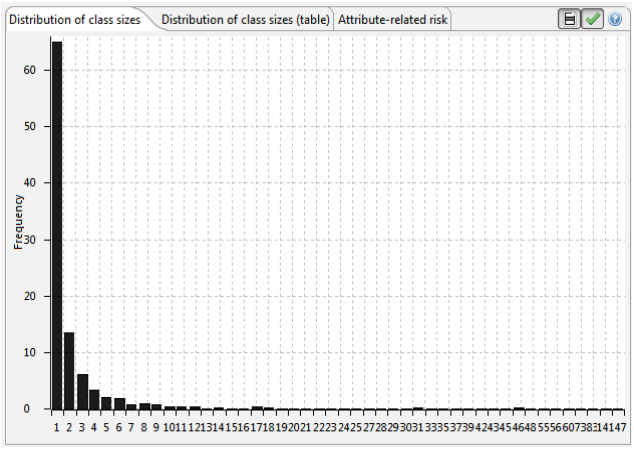

Figure 6: Distribution of class size of Anonymized data D'.

This figure shows the distribution of class after anonymization is performed. This distribution shows that how similar the original data looks after performing 3-anonymization. It is necessary that the anonymized data of every equivalence class looks similar so that no adversary can be understood by any person.

Risk estimation

Risk estimates | Population uniqueness (Danka | Population uniqueness (all mo | Population | Selected attributes

Estimate	Value [%]
Lowest re-identification risk	5.55556
Individuals affected by lowest risk	0.31623
Average re-identification risk	80.34083
Highest re-identification risk	100.00000
Individuals affected by highest risk	69.32537
Entries unique within sample	69.32537
Entries unique within population	22.16821
Population model	PITMAN
Quasi-identifiers	age, education, marital-status, native-country, occupation,...

Figure 7: Risk estimation of original data D.

This figure shows the identification risk of any individual record that the adversary can easily find any person record. Here the identification risk of original data is shown.

Risk estimates | Population uniqueness (Dankar et al.) | Population uniqueness (all models) | Population

Estimate	Value [%]
Lowest re-identification risk	0.14771
Individuals affected by lowest risk	11.89389
Average re-identification risk	2.82853
Highest re-identification risk	20.00000
Individuals affected by highest risk	1.66901
Entries unique within sample	0.00000
Entries unique within population	0.00000
Population model	DANKAR
Quasi-identifiers	age, education, marital-status, native-country, occupation,..

Figure 8: Risk estimation of Anonymized data D'.

This shows the identification risk of 3- Anonymized data. and the identification risk factor of 3-Anonymized data is much less than anonymized data.

Population uniqueness

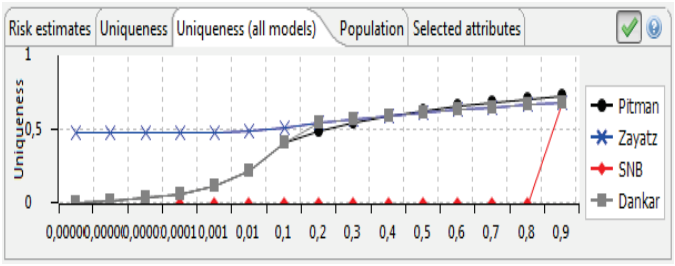

Figure 9: Analysis of Population uniqueness of original data D

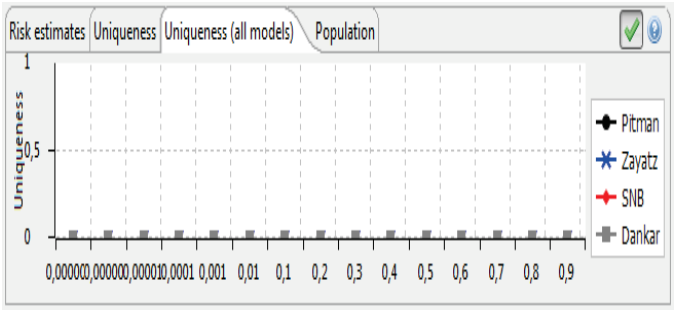

Figure 10: Analysis of Population uniqueness of Anonymized data D'.

Population uniqueness: Population uniqueness means that the total records which are unique in the data or sample.

Conclusion: Heuristic based approach is applied for privacy preserving static data mining. Proposed approach is tried to keep the relationship between the sensitive data and anonymized data. So that no one can forge the sensitive information from the data set. Here all attributes are independent attribute except the sensitive attributes. Here the risk analysis of anonymized data is decreased. Here we have only used l-diversity method. Future work may propose extending our algorithm using t-closeness

References:

1. 7. Dasseni E., Verykios V., Elmagarmid A. and Bertino E., Hiding association rules by using confidence and support, Proceedings of the 4th international workshop on information hiding, pp. 369-383,2001.
2. 8. Domingo-Ferrer J. and Mateo-Sanz J., Practical data-oriented micro aggregation for statistical disclosure control, IEEE transaction on knowledge and data engineering, pp. 189- 201, 2002.
3. 9. Dutta H., Kargupta H., Datta S. and Sivakumar K., Analysis of privacy preserving random perturbation techniques: further explorations, Proceedings of the workshop on privacy in the electronic society (in association with the 1 0•h ACM conference on computer and communications security), 2003.
4. 40 Yumi A., Privacy protection against ubiquitous marketing, SICE Annual Conference, pp. 1434-1436, 2008.
5. 41. Xiaolin Z. and Hongjing B., Research on privacy preserving classification data mining based on random perturbation,International Conference on lnfonnation, Networking and Automation (IC!NA), Vol. I, No. I, pp. 173-178, 2010.
6. 42. Haisheng L., Study of privacy preserving data mining, 3rd Internat ional Symposium on Intelligent Information Technology and Security Informatics, pp. 700-703, 2010.
7. 12. LeFevre K., DeWitt D. and Ramakrishnan R., Incognito: Efficient full domain k-anonymity., Proceedings of the ACM SIGMOD international conference on management of data, pp. 49-60, 2005.

8. 13. Li N., Li T. and Venkatasubramanian S., !-closeness: Privacy beyond k-anonymity and !-diversity, Proceedings oLthe IEEE 23rd International conference on data engineering, 2007,
9. 21. P.Samarati ,"Protecting respondents' identities in microdata release," In IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 13, issue 6, pp 1010-1027, 2001.
10. S. Fienberg and J. McIntyre, "Data Swapping: Variations on a Theme by Dalenius and Reiss," Technical Report, National Institute of Statistical Sciences, 2003.
11. 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. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
12. Christy Thomas, Diya Thomas,Dept of Computer Science & Engineering Rajagiri School of Engineering and Technology, Kochi," An Enhanced Method for Privacy Preservation in Data Publishing"IEEE 2013.
13. Jian Wang,Yongcheng Luo, Yan Zhao,Jiajin Le College of Information Science and Technology, Donghua University Shanghai, China "A Survey on Privacy Preserving Data Mining" 2009 IEEE.
14. P.Samarati ,"Protecting respondents' identities in microdata release," In IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 13, issue 6, pp 1010-1027, 2001.
15. Machanavaijhala A., Gehrke J., Kifer D. and Venkitasubramaniam M., !-diversity: Privacy beyond k-anonymity, Proceedings of the 22nd international conference on data engineering, 2006.
16. 8. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. "*l*-diversity: Privacy beyond kanonymity". In ICDE, 2006.
17. Y. Lindell and B. Pinkas , "Privacy preserving data mining, Journal of Cryptology," vol. 15, issue 3, pp 177-206, 2002.
18. X. Xiao and Y. Tao, "Personalized Privacy Preservation," ACM SIGMOD Conference, 2006.
19. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi and A. W. C. Fu, "Utility Based Anonymization using Local Recoding," ACM KDD Conference, 2006.
20. ARX manual (http://arx.deidentifier.org /anonymization-tool/)
21. UCI Machine Learning Repository
22. (http://archive.ics.uci.edu/ml/datasets)

Aniket patel
CE Department, Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat, India

Nisha Khurana
IT Department, Silver Oak College of Engineering & Technology
Ahmedabad, Gujarat, India