CHAPTER 1

INTRODUCTION TO PRIVACY-PRESERVING IN DATA MINING

1.1 DATA MINING

Data mining uses various data analysis tools to discover patterns/relationships in data to make valid predictions (Sumathi et al 2006; Zaïane 2004). Data Mining, also called Knowledge Discovery in Databases (KDD), refers to nontrivial extraction of previously unknown/useful information from databases. Though data mining and KDD are treated as synonyms, data mining is part of knowledge discovery process (Venugopal et al 2009; Cavoukian et al 1998).

Data mining is an iterative/interactive discovering something innovative. Data mining differs from On-Line Analytical Processing (OLAP) as instead of verifying hypothetical patterns, it uses data to uncover patterns. It is an inductive process (Zaïane 2004). Data mining uses advances in artificial intelligence and statistics. Both disciplines were working on pattern recognition and classification problems. Both contributed to the understanding and application of neural nets and decision trees.

Data mining are automated techniques to extract buried/unknown information from large databases. Data mining is used for four purposes:

i. to improve customer acquisition/retention;

ii. to identify internal inefficiencies and then revamp operations;

iii. to reduce fraud, and

iv. to map unexplored internet terrain.

The primary tools used in data mining include neural networks (NN), decision trees, rule induction, and data visualization (Singh and Yadav, et al 2013).

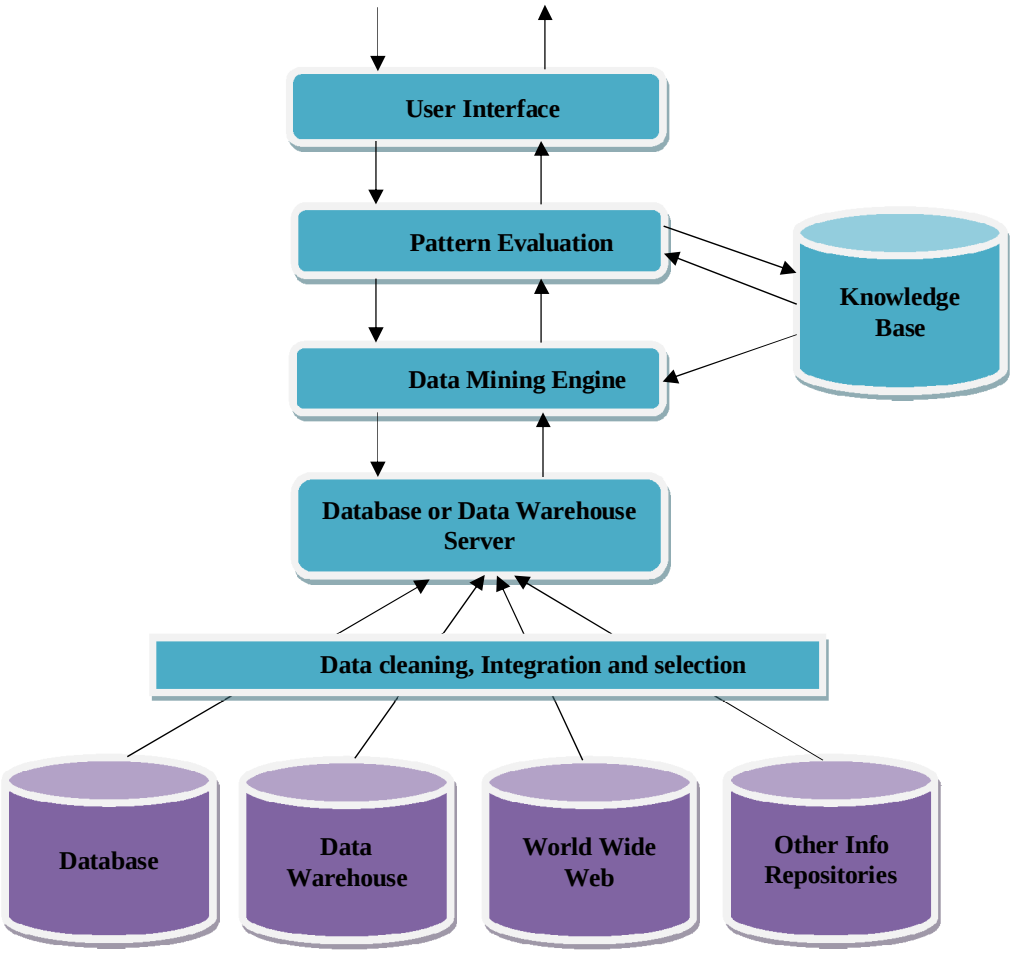

Figure 1.1 Data mining functionalities

Data mining includes three basic steps:

The first processing step is data preparation, often called "scrubbing data." Data is selected, cleansed, and preprocessed under a domain experts guidance and knowledge. Second, data mining algorithm is processes prepared data, compressing and transforming it to ensure easy identification of valuable information. The third phase is data analysis where data mining output is evaluated to check discovery of any additional domain knowledge and determine the relative importance of mining algorithms generated facts.

1.2 PRIVACY ISSUES RELATED TO DATA MINING

Privacy is a thought to be a social/cultural concept. But, with ubiquity of computers and emergence of the Web, privacy is now a digital issue (Han et al 2006). Due to Web revolution and data mining emergence, privacy concerns pose technical challenges different from those before the information era. In an information technology era, privacy refers to users right to conceal their personal information and have some control over use of personal information disclosure. Hence, the privacy concept is more complex than understood. Specifically, privacy preservation definition is still unclear in data mining.

Data mining was developed to provide tools to automatically/ intelligently transform large data knowledge relevant to users. The extracted knowledge, expressed as association rules, decision trees or clusters, permits locating patterns/regularities buried in data but meant to facilitate decision making. This knowledge discovery process returns sensitive information about individuals, compromising their right to privacy. Data mining techniques also reveal critical information about business, compromising free competition, and so disclosures of confidential/personal information should be prevented in addition to knowledge considered sensitive in a given context.

Hence, research was devoted to addressing privacy preservation in data mining resulting in many data mining techniques which included privacy protection mechanisms based on various approaches. Various sanitization techniques were proposed to hide sensitive items/patterns based on removing reserved information/inserting noise in data. Privacy preserving classification methods prevent miners from constructing classifiers capable of predicting sensitive data. Also, recently proposed privacy preserving clustering techniques distort sensitive numerical attributes but preserve general features for cluster analysis (Bertino et al 2005).

Some ways in which privacy concerns raised by data mining are as follows (Oliveira et al 2004),

i. The implicit patterns involving information about persons that can be derived from data in the data-mining process vs. the explicit nature of the personal data (in records) extracted in traditional database retrieval techniques.

ii. The use of (possibly) a single database (or data warehouse') to extract information about persons vs. the use of multiple databases to exchange and retrieve such information.

iii. The use of 'open-ended' queries to discover information on relationships and associations about individuals and groups of individuals vs. (traditional) specific queries to retrieve information about relationships and associations that are already known to exist.

iv. The non-predictive aspect of information about persons gained from data mining vs. the generally predictive aspect of information retrieved from traditional database techniques.

v. The public nature of much of the information about persons that is extracted through the data mining process vs. the private or intimate nature of the information about persons retrieved and exchanged in traditional database-exchange techniques.

vi. The ability to construct new groups or categories of persons based on patterns of information derived from data mining vs. the mere extraction of information about individuals themselves from personal data accessible to traditional techniques of database retrieval.

1.3 PRIVACY PRESERVING DATA MINING (PPDM)

Privacy-Preserving Data Mining (PPDM) is a data mining and statistical databases innovative field where data mining algorithms are analyzed for side-effects in data privacy. It is also called privacy-enhanced/privacy-sensitive data mining dealing with getting valid data mining results without learning underlying data values. This reveals how many different methods and techniques can be used in a PPDM context from a technical perspective.

PPDM has emerged to protect the privacy of sensitive data and also give valid data mining results. Figure 1.2 shows a distributed PPDM scenario which can achieve reasonable privacy and good accuracy. Often a trade-off between privacy and accuracy are needs to be made. On the one hand, privacy requires that the original data records must be fully obfuscated before data mining analysis. On the other hand, accuracy needs that the “patterns” in the original data should be mined out in spite of the perturbation (Likun Liu et al 2012).

There are two major methods in PPDM

- First by using cryptographic representation and
- The other is by using heuristic algorithms which ensures that sensitive data is not revealed.

Most of the current industry requires that these data can be secured during transmission and also when the data is present in the data warehouse (Kumar et al 2013). Originally PPDM extended the traditional data mining techniques to work with data hiding sensitive information, but the major issue was how to modify data and how to recover data mining results from that modified data. The goals of a PPDM algorithm include:

i. Prevent the discovery of sensible information.
ii. Being uncompromised to access and to use the non-sensitive data.
iii. Being usable on large amounts of data.
iv. Must have less exponential computational complexity.

Generally PPDM techniques are based on cryptography, data mining and information hiding. Knowledge models such as decision trees are used on sanitized data. This approach includes an advantage of its efficiency in handling large datasets volumes (Mandapati et al 2013). There are two fundamental problems of PPDM: privacy-preserving data collection and mining a data set partitioned across several private enterprises (John and Deepajothi 2013). The aim of these algorithms is to retrieve knowledge from data warehouse while preserving the confidentiality of data.

In recent years, many PPDM methods have been developed but there is no standardization in these approaches. To achieve optimized results while preserving the privacy of the data subjects efficiently, five dimensions need to be considered and listed below:

(1) The distribution of the basic data

(2) The modification of the basic data

(3) Mining method being used

(4) If basic data or rules are to be hidden and

(5) Additional methods for privacy preservation used.

This shows that from a technical viewpoint many different methods and procedures in the perspective of PPDM that can be used (Kiran et al 2012).

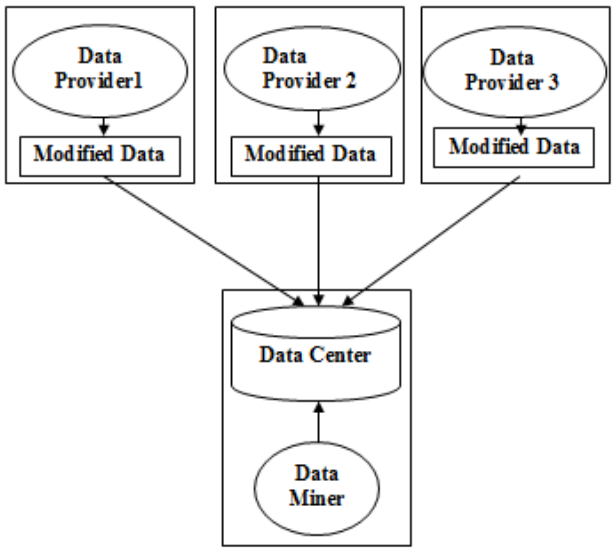

Figure 1.2 PPDM based on data publishing scenario

Data mining needs correct input for meaningful results, but privacy concerns influence users to provide wrong information. To preserve client privacy in data mining procedures, various random perturbation of data records based techniques were proposed. Randomization/Distortion are two methods that preserve privacy. Randomization modifies transactions through replacing some items with non-existing items and also through the addition of fake items to ensure privacy preservation. Distortion operates on a transaction database through probabilistically changing items in every transaction (Shrivastava et al 2011).

1.3.1 Models of PPDM

i. Trust Third Party Model

The security standard assumes we have a trusted third party to which all data is given. The third party performs computation and delivers results and except for this party, nobody learns anything inferable from own input/ results. Secure protocols aim to reach this privacy preservation level without finding a third party everyone trusts.

ii. Semi-honest Model

In this, all parties follow protocol rules using correct input, but when the protocol is free it uses anything it sees during protocol execution to compromise security.

iii. Malicious Model

In malicious model, participants have no restrictions. Any party is free to indulge in any action. Usually, it is difficult to develop efficient protocols valid under a malicious model.

iv. Other Models - Incentive Compatibility

Though semi-honest and malicious models are well researched, other models outside purview of cryptography are also possible. An example is incentive compatibility. A protocol is incentive compatible when a cheating party is either caught/suffers an economic loss. Under the rational economics model, this ensures that parties have no advantage by cheating. Of course, this fails in an irrational model. (Ge X et al 2010).

1.4 NEED FOR PPDM

Data mining techniques were developed to extracts knowledge to support various domains like weather forecasting, marketing, medical diagnosis and national security. But it is still challenging to mine specific data without violating data owners ’privacy. For instance, mining patients ’private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns increase. Commercial issues are also linked to the privacy issue. Most organizations collect information about individuals for specific needs. Frequently different units in an organization may find it necessary to share information. In such cases, each organization/unit must ensure that individual privacy is not violated or sensitive business information revealed. To avoid these types of violations, there is a need of various data mining algorithm for privacy preserving (Nayak et al 2011).

1.5 APPLICATIONS OF PPDM

Privacy is an important issue in many data mining applications and it deals with some application fields such as

- Health care,
- Security,
- Financial and
- Other types of sensitive applications (Kamakshi et al 2010).

1.6 PPDM TECHNIQUES

PPDM is divided into two groups: data hiding and rule hiding. The objective of data hiding is to transform data or design new computation protocols to ensure that private data remains private during or after data mining; when underlying data patterns/models can be discovered. Techniques like multiplicative perturbation, additive perturbation and secure multi-party computation fall in this category. Rule hiding, on the other hand, transforms database so that sensitive rules are masked still allowing underlying patterns to be discovered.

Most privacy computation methods use some form of transformation on data to ensure privacy preservation. Such methods reduce representation granularity to reduce privacy. This results in some data management/mining algorithms loss of effectiveness; a natural trade-off between information loss and privacy. Usually such methods reduce representation granularity to reduce privacy. This reduction leads to some loss in data management/mining algorithms effectiveness - a trade-off between information loss and privacy. The most common techniques are Randomization techniques, group based anonymization and distributed PPDM

1.6.1 Randomization Techniques

Randomization technique is the process of perturbing the input data to distributed data mining algorithms so that the data values of individual

entities are protected from revealing. Several randomization techniques has been identified in PPDM algorithms by including

- Adding random numbers,
- Generating random vectors and
- Random permutation of a sequence.

Data can be perturbed in two manners: the value class membership and value distortion. The value class membership is a method that values of an attribute are divided into intervals and the interval in which a value lies is returned instead of the original value. The value distortion method works by adding a random value $y_i$to each value $x_i$of an attribute. Then, the original data distribution is reconstructed by the Bayesian approach, i.e., iterating

$$f_X^{(j)}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_y(w_i - a)\, f_X^{j-1}(a)}{\int_{-\infty}^{\infty} f_y(w_i - z)\, f_X^{j-1}(z)dz} \qquad (1.1)$$

until $f_X^{(j)}$is statistically the same as the original distribution of X, where $X = (x_1, x_2, \ldots . x_n)$ is the original variable, $Y = (y_1, y_2, \ldots . y_n)$ is an random variable obeying a uniform distribution between [-u, u], $f_Y(a)$ stands for the density function of Y, $w_i = x_i + y_i$ for i =1, 2, …, n, and for $f_x^{(0)}$is a uniform distribution. Given a sufficiently large number of samples, $f_x^{(j)}$ can be expected to be very close to the real density function f of X after sufficient iterations. Based on the reconstructed distribution, decision trees can be induced (Xu et al 2011).

Randomization is a technique easily implemented during data collection as noise added to a record is independent of other data records behaviour. This is a weakness as outlier records are difficult to mask. Clearly,

where privacy-preservation is not required at data-collection time, a technique where inaccuracy depends on behaviour of the locality of that record is necessary. Another randomization framework weakness is that it does not consider the chances that available records can identify the that record's owners. It was shown that use of publicly available records leads to privacy being compromised in high-dimensional cases (Aggarwal 2007). This holds good for outlier records easily distinguished from others in their locality. Hence, a broad privacy transformation approach is constructing anonymous records groups that are transformed in group-specifically.

1.6.2 Group based Anonymization

There are many privacy models associated with group based anonymization such as k-anonymity (Sweeny 2002), l-diversity (Machanavajjhala et al 2006), t-closeness (Li and Li 2007), (k, e)-anonymity (Wong et al 2006), Injector (Li and Li 2008) and m-confidentiality (Wong et al 2007). The group based anonymization methods such as K-anonymity and l-diversity are satisfied by using a generalization that replaces a value with a less specific, more general value.

1.6.2.1 K-anonymity Technique and L-diversity Technique

In K-anonymity technique ,each record in an anonymized table must be indistinguishable with at least k-1 other record within the dataset, with respect to a set of QI attributes. In particular, a table is K-anonymous only if the Quasi Identifier (QI) attributes values of each record are identical in those of at least k-1 other records. To achieve the K-anonymity requirement, generalization or suppression concept can be used (Keyvanpour et al 2011).

k-anonymity is attractive process due to its simple definition and numerous algorithms available for anonymization. But still it is vulnerable to many types of attacks specially when attacker has background knowledge. In Homogeneity attacks, all sensitive attribute values in a group of k records are similar. So even if data is k-anonymized, sensitive attribute values for that group of k records are easily predictable. In background knowledge attack, adversaries use an association between one/more quasi-identifier attributes with sensitive attribute to narrow down sensitive field values further.

So, while k-anonymity prevents record identification, it is not always effective in preventing sensitive values inference of a record's attributes. Hence, l-diversity was proposed as it both maintains minimum group size of k, and maintains sensitive attributes diversity. A data set satisfies l-diversity if, for every records group sharing a key attributes combination, there is minimum of l "well represented" values for every confidential attribute.

L-diversity technique has been proposed to solve the Homogeneity attack of K-anonymity technique that emphasizes not only on saving the minimum size of K group but it also considers saving the variety of the sensitive attributes of each group. Here every anonymized group holds at least l well-represented values for each sensitive attribute. Though, L-diversity technique includes shortcoming in it; that is insufficient to prevent attribute disclosure such as similarity attack (Machanavajjhala et al 2006).

1.6.2.2 t-closeness

The t-closeness is enhancement on l-diversity concept. A characteristic of l-diversity model is that it treats values of an attribute similarly, irrespective of its data distribution. This is rare for real data sets, as attribute values may be skewed making it tough to create feasible l-diverse

representations. Usually, adversaries use global distribution background knowledge to infer sensitive data values. Further, not all attribute values are equally sensitive. For instance, an attribute corresponding to a disease may be more sensitive with positive values and not when it is negative. A t-closeness model proposed uses property that distance between distribution of a sensitive attribute in an anonymized group should not differ from global distribution by more than a threshold t (Li &Li 2007). The Earth Mover distance metric quantifies distance between both distributions. Also, t-closeness approach is more effective than other PPDM methods for numeric attributes.

1.6.3 Distributed Privacy Preservation

The goal in privacy-preserving data mining distributed methods is that to allow a computation of useful aggregate statistics over entire data set without compromising individual data sets privacy within different participants. Hence, participants may desire to collaborate to get aggregate results, but be unable to trust each other fully regarding distribution of own data sets. Data sets for this purpose are either horizontally or vertically partitioned. In the former, individual records are spread out across multiple entities where each has same attributes. In vertical partitioning, individual entities have different attributes of same set of records. Both partitioning attitudes offer different challenges to distributed privacy-preserving data mining (Nayak et al 2011).

1.7 DATA PERTURBATION

Data perturbation techniques are popular models for PPDM and are used in applications where data owners wanting to participate in cooperative mining simultaneously want to prevent privacy-sensitive information leakage in published datasets. Examples include micro data publishing for research or data outsourcing to a third data mining service provider.

Statistical databases Inference control also called Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL), protects statistical data so that they are publicly released/mined without release of private information that links specific individuals/entities. SDC techniques protection entails some data modification, an intermediate option between no modification (maximum utility, without disclosure protection) and data encryption (maximum protection without utility for user sans clearance). SDC challenge is modifying data to ensure sufficient protection while keeping information loss (accuracy loss sought by database users) to a minimum. Micro data protection methods generate protected microdata set V' by either masking original data, i.e. generating V' a modified original microdata set V version or by generating synthetic data V' preserving some statistical original data V properties. Masking methods are divided in two categories: Perturbative Masking Methods and Non-Perturbative Masking Methods.

1.7.1 Perturbative Masking Methods

Perturbative methods release entire microdata set, though it is the perturbed values and not exact values that are released. All perturbative methods are not designed for continuous data; this distinction is addressed further in this paper for each method. Most perturbative methods reviewed below (rank swapping, additive noise, micro-aggregation and post-randomization) are special matrix masking cases. If original microdata set is X, then masked microdata set Z is computed as

$$Z = AXB + C \qquad (1.2)$$

where A is a record-transforming mask, B an attribute-transforming mask and C a displacing mask (noise)(Duncan and Pearson.1991).

Correlated noise addition preserves additionally allowing preservation of correlation coefficients in the additive noise method. The difference between this and the earlier method is that errors covariance matrix is now proportional to original data covariance matrix. It suits continuous data as there is no assumption on range of possible values for V_i (which can be infinite), added noise is continuous and with mean zero, which suits continuous original data well. Exact matching is impossible with external files. Depending on the noise added, approximate (interval) matching is possible.

Data swapping was earlier presented as an SDC method for databases having categorical attributes (Dalenius & Reiss 1978) alone. The idea being to transform a database by exchanging confidential attributes values among individual records. Records are exchanged ensuring that low-order frequency counts/marginals are maintained. Rank swapping is used for a numerical attribute (Moore 1996). First, values of attribute *Xi* are ranked in ascending order, then every ranked value of *Xi* is swapped with another randomly chosen in a short range (rank of two swapped values cannot differ by more than *p*% of total records, where *p* is input parameter). This algorithm is used independently on every original attribute in original data set.

Micro-aggregation belongs to the perturbative SDC methods family originally designed for continuous numerical data and later extended for categorical data (Domingo-Ferrer et al 2006). Whatever be the data type, micro-aggregation is operationally defined regarding two steps: Partition and Aggregation. During partition, original records set is partitioned into many groups so that records in same group are similar to others; so that number of records in every group is at least k. A partition which meets this requirement on minimal group size is called a k-partition. An aggregation operator (mean

for numerical data) computes a centroid for every group when each record in a group is replaced by a group centroid.

A general perturbative method to mask microdata records against re-identification is PRAM (Post-Randomization Method) for categorical variables (Gouweleeuw et al 1998). PRAM adds random noise to continuous variables where values of categories are changed/not changed depending on a prescribed probability matrix and stochastic process based on random multinomial draw outcome. The prescribed probability matrix is developed so as to preserve original variable's anticipated marginal frequencies to reduce information loss. Use of a more deterministic approach in perturbation maintains exact marginal distributions. PRAM introduces misclassification into microdata which cause perturbed records to fail edit constraints. Hence, a PRAM procedure which will simultaneously consider edit constraints and ensure that resulting perturbed microdata satisfy all edits should be developed. Though perturbed microdata users know that certain records variables are misclassified, it is inadvisable to release microdata with failed edit records as this damages data utility. Additionally, an inconsistent and illogical record immediately aims at the perturbed record attempting to unmask it. This is the case when microdata contain hierarchical data (households and persons) and unperturbed variables can identify perturbed variables and original content.

1.7.2 Non-Perturbative Masking Methods

Original data is not modified, but some data is suppressed and/or some details removed; Non-perturbative techniques result in protected micro data by eliminating details from original micro data. Some nonperturbative techniques are Local suppression, Sampling, Global recoding and Generalization.

Sampling methods suit categorical microdata, but they should be combined with other masking methods for continuous microdata. This is because sampling alone leaves a continuous attribute *Vi* unperturbed for all records in *S*. Thus, if attribute *Vi* is present in external administrative public file, unique matches with published sample are likely: given a continuous attribute *Vi* and two respondents *o*1 and *o*2, it is unlikely that *Vi* will take same value for both *o*1 and *o*2 unless *o*1 = *o*2 (this is true even if *Vi w*as truncated to represent it digitally).

Recoding is a variable's collapsing categories, suppression is replacement of a record's value by a missing value, and perturbation is replacement of one value by another. Protecting microdata sets by one of the above measures results in information loss., Certain individual attributes values are suppressed to increase the set of records agreeing on a key values combination in local suppression. If a continuous attribute *Vi*is part of a key attributes set, then each key values combination is unique. As it is not sensible to systematically suppress values of *Vi*, it is concluded that local suppression is categorical attributes oriented.

1.7.3 Multiplicative Perturbation for PPDM

Multiplicative perturbation algorithms improve data privacy while maintaining desired data utility level by selective preservation of the mining task, modelling specific information during data perturbation. By preserving task and model specific information, a "transformation-invariant data mining models" set is applied to perturbed data directly, achieving needed accuracy. Usually, a multiplicative perturbation algorithm finds multiple data transformations which preserve required data utility. Hence, the next challenge is locating a good transformation that provides a satisfactory privacy guarantee.

This category includes three particular perturbation techniques types: Rotation Perturbation, Projection Perturbation, and Geometric Perturbation. Compared to other multi-dimensional data perturbation methods, perturbations reveal unique properties for privacy preserving data classification/data clustering. They preserve (approximately preserve) distance/inner product, important for classification and clustering models. So, perturbed data based classification/clustering mining models show similar accuracy to those based on original data through multiplicative data perturbation. The challenge for multiplicative data perturbations in knowing how to maximize desired data privacy as other data perturbation techniques seek a better trade-off between data utility level and accuracy preserved and level of data privacy guaranteed (Chen & Liu 2008).

1.8 K-ANONYMITY

K-anonymization mainly aims at preventing sensitive information about individuals being identified or inferred from the dataset. In case of *k*-anonymity, the system masks the values of some potentially identifying attributes, called “quasi-identifier”. According to this principle each record in a relational table T needs to have the same value over quasi-identifiers with at least *k*-1 other records in *T*. It is considered as a better protection than exposing all the information in the dataset (Pasierb et al).

1.8.1 Existing Techniques

The k-anonymity requirement is enforced through generalization, where real values are replaced with “less specific but semantically consistent values” (Sweeney 2002). there are various ways to generalize values in a domain. Numeric values are generalized into intervals (12−19), and categorical values are into a set of distinct values ({USA, Canada}) or a single value representing such a set (North-America).

Differing generalization strategies were proposed. In (LeFevre et al 2005; Sweeney 2002), a non-overlapping generalization-hierarchy is defined first for every quasi identifier attribute. An algorithm then attempts to locate an optimal (or good) solution which allowed by such generalization hierarchies. In these schemes, when a lower level domain must be generalized to a higher level, all lower domain values are generalized to higher domain. This restriction is a major drawback as it leads to relatively high data distortion during unnecessary generalization. Algorithms allow values from varied domain levels to combine to represent a generalization (Fung, et al 2005). Though this leads to a more flexible generalization, they are still limited by imposed generalization hierarchies.

Schemes that don't rely on generalization hierarchies were proposed. For instance, LeFevre et al (2006) transforms k-anonymity problem to a partitioning problem. Specifically, the approach includes the following 2 steps. The first is to find partitioning of d-dimensional space, where d is number of quasi-identifier attributes, so that each partition has k records minimum. Then records in each partition are generalized so that all share same quasi-identifier value. Though efficient, these approaches are also disadvantageous requiring a total order for each attribute domain making it impractical in cases involving categorical data with no order.

1.8.2 Advantages of K-anonymity

K-anonymity model is simple, intuitive, and well-understood. It appeals to non-expert, the model's end client. This protects respondents' identities while releasing truthful information. The k-anonymity model defines process output privacy and not of process itself in contrast to the majority of privacy models suggested earlier. It is in this sense of privacy that interests clients.

Though computing k-anonymous table is hard (Meyerson and Williams 2004), it is easy to validate as an outcome is k-anonymous. Hence, this assures non-expert data owners that model are used properly.

1.8.3 Disadvantages of K-anonymity

This technique consists of some limitations.

- It is very difficult for a database owner to determine which of the attributes are or are not available in external tables.
- It considers a certain type of attack such as linkage attack and cannot preserve sufficiently the sensitive attributes against the homogeneity attack which is a similarity of the sensitive attributes values in an anonymized group and background knowledge attack such as awareness about the relationship between sensitive and QI attributes (Mohammad Reza Keyvanpour et al 2011).
- K-anonymity may leak private information. An adversary has strong background knowledge about the sensitive values; adversary may be able to infer non-sensitive values from the sensitive values. These vulnerabilities are both caused by a lack of diversity in the sensitive values (Junichi Sawada et al 2012).
- K-anonymity is difficult to achieve before all data are collected in one trusted place
- Attributes that are not among quasi-identifiers, even if sensitive (e.g., diagnosis), are not suppressed and may get linked to an identity.

1.9 MOTIVATION

Many applications using data mining techniques involve mining subject's private/sensitive information. A way to enable effective data mining while preserving privacy is anonymize data set that includes subjects private information before releasing it for data mining. A way to anonymize data set is by manipulating content so that records adhere to k-anonymity. Two common manipulation techniques to achieve dataset k-anonymity are generalization and suppression. Generalization is replacing a value with a less specific but semantically consistent value, while suppression is not releasing a value totally. Generalization is more commonly used as suppression dramatically reduces data mining results quality when improperly used. But, generalization's drawback is that it requires manually generated domain hierarchy taxonomy for every dataset quasi-identifier to perform k-anonymity.

K-Anonymity is a privacy preserving method for limiting disclosure of private information in data mining. The process of anonymizing a database table typically involves generalizing table entries and, consequently, it incurs loss of relevant information. This motivates the search for anonymization algorithms that achieve the required level of anonymization while incurring a minimal loss of information. The problem of k-anonymization with minimal loss of information is NP-hard.

1.10 OBJECTIVE OF PRESENT INVESTIGATION

K-anonymity techniques are based on the reduction of granularity of representation of data using pseudo-identifiers. Major techniques used for granularity reduction was generalization and suppression. In generalization, the attribute values are converted into a range that reduces the granularity and reduces the risk of identifying individual values. In suppression method,

actual value of the attribute is removed completely. But these two methods introduce loss of some detail which may affect the accuracy.

Finding optimal k-anonymous datasets using generalization or suppression has been proved as a NP-hard problem (Winkler; Atallah et al 1999). So some standard heuristic search techniques such as genetic algorithms, particle swam optimization and ant colony optimization can be used to find optimal datasets Following are the objectives of the research:

- The effect of the anonymization due to k-anonymity on the data mining classifiers is investigated.
- Optimize search for right tradeoff between privacy and information loss using Genetic Algorithm (GA) is proposed.
- Hybrid optimization based on Simulated Annealing (SA) with GA is proposed to preserve the classification accuracy.

1.11 THESIS ORGANIZATION

This chapter presented a detailed overview of Privacy Preserving Data Mining (PPDM). A summary of Data mining and privacy concerns is also presented. A brief description of various techniques used in PPDM is presented. The motivation and the objectives of this research are discussed.

In chapter 2, a review of works available in the literature is included. The chapter deals with reviews covering earlier works related to PPDM Algorithms, modifications of k-anonymity techniques, feature selection in PPDM and optimization techniques used in PPDM.

In chapter 3, the performance of classifier to classify non-anonymized and anonymized datasets is evaluated. Experiments are conducted using Mushroom dataset and Integrated Public Use Microdata Series (IPUMS) dataset.

Chapter 4 explores the effectiveness of genetic optimization for feature selection in PPDM.

In chapter 5, the implementation of the proposed hybrid optimization based on Simulated Annealing (SA) with GA is discussed and the experimental results presented. Chapter 6 concludes the study and the future works are discussed.