ISSN:0975-9646

Comparative Study of Privacy Preservation Techniques in Data Mining

Avruti Srivastava
Department of Computer Engineering,
MCT's Rajiv Gandhi Institute of Technology
Mumbai, Maharashtra

Abstract: **Extraction of knowledgeable information from large quantities of data and its study is crucial to Data mining. The process of data collection and data dissemination may, however, result in an inherent risk of privacy threats. The privacy-preserving data mining (PPDM) has thus become an important issue. Various methods have come into existence. The methods can be grouped into various categories based on their methodology and working, for example Integer Partitioning Based Encryption (IPBE) is an effective cryptography based method for the same. This method provides a greater amount of protection compared to usual cryptographic techniques as it is not a straight forward public key encryption method.**

Keywords— privacy preservation, encryption, data mining, integer based partitioning

I. Introduction

Recent advances in information, communications, data mining, and security technologies have given rise to a new era of research, known as Privacy Preserving Data Mining (PPDM). Several encryption algorithms, incorporating privacy preserving mechanisms, have been developed that allow one to extract relevant knowledge from large amount of data, while hide sensitive data or information from disclosure or inference.

The information found in mining can be sensitive or it can be misused by anyone. Consider a scenario in which two or more parties own their confidential databases wishes to run a data mining algorithm on the union of their database without revealing any private information. For example, separate medical institutes wish to conduct a join medical research while preserving the privacy of their patients. In this condition it is required to protect the sensitive information, but it is also required to enable its use for future research work. Involving parties are realizing that combining their data gives mutual benefit but none of them is willing to reveal its database to other parties. For this reason various privacy preserving techniques are applied with data mining algorithm in order to protect the extraction of sensitive information during the knowledge finding.

Consider another example, the progress of hardware technology has made it easy to store and process large amounts of transactional information. Even simple transactions of everyday life such as using the phone or credit-cards are recorded today in an automated way. A large amount of such information is often specific to individual users. Depending upon the nature of the information, users may not be willing to divulge the individual values of records. In particular, data mining techniques are considered a challenge to privacy preservation due to their natural tendency to use sensitive information about individuals. This has led to a considerable amount of focus on privacy preserving data collection and mining methods.

II. Privacy Issues Related To Data Mining

Data mining is included in day to day operational activities of every organization. During the whole process of data mining (from collection of data to discovery of knowledge) we get the data. These data may contain sensitive information of individuals. This information may expose to several other entities including data collector, users and miners. Disclosure of such information results in breaking the individual privacy. for ex: exposed credit card details of a user will affect its social and economic life. Private information can also be disclosed by linking multiple databases belong to giant data warehouse and accessing web data. A malicious data miner can learn sensitive data values or attributes such as income or disease type of a certain individual through re-identification of record from an exposed data set. The combination of other attributes also provides help to the intruders to identify the sensitive data values. If we remove other attributes then it's not guaranteed to provide the confidentiality of private information. Sufficient supplementary knowledge is also helpful for intruders to identify sensitive data values.

A. Public Awareness

Now a days secondary use of data become very common. Secondary use of data means data is used for some other purpose not for which data is collected initially. The potential misuse of personal information of public is increasing rapidly. The scope of sensitive data is not limited to medical or financial records it may be phone calls made by an individual, buying patterns and many more. No one wants that his/her personal data is sold to any other party without their prior consent. Some individuals become hesitant to share their information which results additional difficulty to obtaining correct information. Public awareness is so much important if private information is shared between different entities. Public awareness about privacy and lack of public trust in organization may introduce additional complexity to data collection. Strong public concern may force government and law forcing agencies to introduce new privacy protecting regulations. For example federal employees being prepossess, on the basis of protected genetic information, according to US Executive Order.

B. Privacy Preserving Data Mining

Due to the tremendous benefit of data mining and high public concern regarding individual data privacy, implementation of privacy preserving data mining has become demand of today's environment. This technique provides individual privacy while at the same time allowing extraction of useful knowledge from data. There are several methods which can be used to enable privacy preserving data mining. Most of the techniques use some form of transformation or modification. These techniques modified the collected data set before its release in an attempt to protect individual records from being re identified. A malicious data miner or intruder even with additional knowledge cannot be certain about the correctness of a re-identification, even when the data set has been modified. Apart from the context of data mining it is important to maintain patterns in data set. High data quality with privacy/security is the major requirement of good privacy preserving techniques.

An effective approach for privacy preserving technique relies on two facts:

- Users are not equally protective of all values in their records. Thus, users may be willing to provide modified values of certain fields by the use of a (publically known) key. This modified value may be generated using a custom code.
- Data mining problems do not necessarily require individual records, but only distributions. Since the distribution is known, it can be used to reconstruct aggregate distributions, i.e. the probability distribution of the data set. In many cases, it can be developed which use the probability distributions rather than individual records.

The increasing ability to track and collect large amounts of data with the use of current hardware technology has led to an interest in the development of data mining algorithms which preserve user privacy. There exists technique that addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining. It helps to retain privacy while accessing the information implicit in the original attributes. The distribution reconstruction process naturally leads to some loss of information which is acceptable in many practical situations. Specifically, there exist algorithms that converge the maximum likelihood estimate of the original distribution based on the perturbed data. For a large amount of data, it provides robust estimates of the original distribution.

III. Privacy Preserving Methods

A. RANDOMIZATION METHOD: The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

1) Additive Perturbation: In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re-designed to work with these data distributions. In Fixed-data perturbation method the data is changed by adding a noise term e to the attribute X resulting in Y, Y=X+e, where e is drawn from some probability distribution. This method is called Additive Data Perturbation (ADP).

2) Multiplicative Perturbation: In this case, the random projection or random rotation techniques are used in order to perturb the records. Privacy preserving data mining is used for secure mining from the data warehouse. Random perturbation technique is a method to convert raw data based on probability which has been discussed. Data distortion is achieved by changing the original data, in which some randomness value is added such that the original data is difficult to ascertain, while preserving global feature of a record. In Multiplicative Data Perturbation (MDP) the value of e is multiplied with X to get Y the perturbed value, Y=Xe, where e has mean of 1.0 and a specified variance.

In randomization perturbation approach the privacy of the data can be protected by perturbing sensitive data with randomization algorithms before releasing it to the data miner. The perturbed data version is then used to mine patterns and models. The algorithm is so chosen that combined properties of the data can be recovered with adequate accuracy while individual entries are considerably distorted. In this method privacy of confidential data can be obtained by adding small noise component which is obtained from the probability distribution .In a set of data records denoted by

X = {x1 . . . x N }. For record x I ЄX, a noise component is added which is drawn from the probability distribution.

Commonly used distributions are the uniform distribution over an interval [−α, α] and Gaussian distribution with mean μ = 0 and standard deviation σ. These noise components are drawn independently, and are denoted y1 . . . yN. Thus, the new set of distorted records is denoted by x1 + y1.... x N + y N. It is denoted by this new set of records z1 . . . z N. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data.

B. K-ANONYMITY: - The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, it reduces the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. An important method for privacy de-identification is the method of k-anonymity.

The motivating factor behind the k anonymity technique is that many attributes in the data can often be considered pseudo-identifiers which can be used in conjunction with public records in order to uniquely identify the records. For example, if the identifications from the records are removed, attributes such as the birth date and zip-code can be used in order to uniquely identify the identities of the underlying records. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least (k − 1) other records.

Age	Name	Weight
35	ABC	65
23	DEF	55
44	XYZ	75

(a) Original Data

Age	Name	Weight
[30,40]	ABC	[60,70]
[20,30]	DEF	[50,60]
[40,50]	XYZ	[70,80]

(b) K-Anonymous Data

C. CRYPTOGRAPHIC TECHNIQUES:- In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores.

This requires secure and cryptographic protocols for sharing the information across the different parties.

Cryptography, the science of communication and computing in the presence of a malicious adversary extends from the traditional tasks of encryption and authentication .In an ideal situation, in addition to the original parties there is also a third party called “trusted party”. All parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. The protocol that is run in order to compute the function does not leak any unnecessary information. Sometimes there are limited leaks of information that are not dangerous. This process requires high level of trust.

D. CRYPTOGRAPHIC APPROACH:- In this novel framework the total process is divided into three components the customer, mediator and a group of service data providers. Initially there is no communication between customer and data provider. When the client sends a query, the mediator send the information to all the data holders and through exchange of acknowledgements, the mediator establishes the connection with data providers. Also, we have a cryptographic approach i.e. Elliptic Curve Cryptography (ECC).

E. ECC APPROACH:- Elliptic Curve Cryptography (ECC) is an attractive alternative to conventional public key cryptography. ECC is an ideal candidate for implementation on constrained devices where the major computational resources i.e. speed, memory is limited and low-power wireless communication protocols are employed.

That is because it attains the same security levels with traditional cryptosystems using smaller parameter sizes. Moreover, in several application areas such as person identification and e-Voting, it is frequently required of entities to prove knowledge of some fact without revealing this knowledge. Such proofs of knowledge are called Zero Knowledge Interactive Proofs (ZKIP) and involve interactions between two communicating parties, the Prover and the Verifier. In a ZKIP, the Prover demonstrates the possession of some information to the Verifier without disclosing it.

Generally, a zero-knowledge protocol allows a proof of the truth of an assertion, while conveying no information whatsoever about the assertion itself other than its actual truth. Usually, such a protocol involves two entities, a prover and a verifier. A zero-knowledge proof allows the prover to demonstrate knowledge of a secret while revealing no information whatsoever of use to the verifier in conveying this demonstration of knowledge to others.

The zero-knowledge protocols are instances of interactive proof systems and non-interactive proof systems. In the first category, a prover and a verifier exchange multiple messages challenges and responses), typically dependent on random numbers which they ma y keep secret whereas in the second the prover sends only one message. In both systems the prover's objective is to convince the verifier about the truth of an assertion, e.g. the claimed knowledge of a secret. The verifier either accepts or rejects the proof. A zero-knowledge proof must obey the properties of completeness and soundness. A proof is complete, if given an honest prover and an honest verifier, the protocol succeeds with vast probability and sound if the probability of a dishonest prover to complete the proof successfully is negligible. A typical example of zero-knowledge proof is known as Alibaba's cave problem. In this story, Annie has uncovered the secret word used to open a magic door in a cave.

John waits outside the cave as Annie goes in. Annie randomly takes either path A or B inside the cave John enters the cave and shouts the name of the path he wants her to use to return either A or B, chosen at random Annie does that using the secret word if needed to open the magic door. The above steps are repeated n times until John is convinced that Annie knows the secret word. Now, suppose that Annie does not know the secret word. Since, John chooses path A or B at random, Annie has a 1/2 chance of cheating at one round. If the above steps are repeated for many rounds, Annie's chance of successfully anticipating all of John's requests would become vanishingly small. Thus, if Annie reliably appears at the exit John names, he can conclude that she is very likely to know the secret word.

The scheme steps are now described.

(1) John waits outside the cave as Annie goes in.

(2) Annie at random takes either path A or B inside the cave.
(3) John enters the cave and shouts the name of the path he wants her to use to return either A or B, selected at random.
(4) Annie does that using the secret word if needed to open the magic door.
(5) The above steps are repetitive n times until John are convinced that Annie knows the secret word.

IV. The *K*-Anonimity Framework

In many applications, the data records are made available by simply removing key identifiers such as the name and social-security numbers from personal records. However, other kinds of attributes (known as pseudo-identifiers) can be used in order to accurately identify the records. For example, attributes such as age, zip-code and sex are available in public records such as census rolls.

When these attributes are also available in a given data set, they can be used to infer the identity of the corresponding individual. A combination of these attributes can be very powerful, since they can be used to narrow down the possibilities to a small number of individuals.

In *k*-anonymity techniques, we reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as *generalization* and *suppression*. In the method of *generalization*, the attribute values are generalized to a range in order to reduce the granularity of representation.

For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. In the method of *suppression*, the value of the attribute is removed completely.

It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

In order to reduce the risk of identification, the *k*-anonymity approach requires that every tuple in the table be in distinguishability related to no fewer than *k* respondents. This can be formalized as follows:

Definition: *Each release of the data must be such that every combination of values of quasi-identifiers can be indistinguishably matched to at least k respondents.*

The approach uses domain generalization hierarchies of the quasi-identifiers in order to build *k*-anonymous tables. The concept of *k*-minimal generalization limits the level of generalization for maintaining as much data precision as possible for a given level of anonymity. Subsequently, the topic of *k*-anonymity has been widely researched.

The approach assumes an ordering among the attributes. The values of the attributes are discretized into intervals (quantitative attributes) or grouped into different sets of values (categorical attributes). Each such grouping is an *item*. For a given attribute, the corresponding items are also ordered.

An index is created using these attribute-interval pairs (or items) and a set enumeration tree is constructed on these attribute-interval pairs. This set enumeration tree is a systematic enumeration of all possible generalizations with the use of these groupings. The root of the node is the null node, and every successive level of the tree is constructed by appending one item which is lexicographically larger than all the items at that node of the tree. The number of possible nodes in the tree increases exponentially with the data dimensionality. Therefore, it is not possible to build the entire tree even for modest values of *n*. However, the k-Optimize algorithm can use a number of pruning strategies to good effect. In particular, a node of the tree can be pruned when it is determined that no descendent of it could be optimal. This can be done by computing a bound on the quality of all descendants of that node, and comparing it to the quality of the current best solution obtained during the traversal process. A branch and bound technique can be used to successively improve the quality of the solution during the traversal process. Eventually, it is possible to terminate the algorithm at a maximum computational time, and use the current solution at that point, which is often quite good, but may not be optimal.

The Incognito method uses a bottom-up breadth-first search of the domain generalization hierarchy, in which it generates all the possible minimal k-anonymous tables for a given private table. First, it checks *k*-anonymity for each single attribute, and removes all those generalizations which do not satisfy *k*-anonymity. Then, it computes generalizations in pairs, again pruning those pairs which do not satisfy the *k*-anonymity constraints. In general, the Incognito algorithm computes (i + 1)-dimensional generalization candidates from the i-dimensional generalizations, and removes all those generalizations which do not satisfy the *k*-anonymity constraint. This approach is continued until, no further candidates can be constructed, or all possible dimensions have been exhausted.

V. Secure Multi Party Communication

Defining privacy: The common definition of privacy in the cryptographic community limits the information that is leaked by the distributed computation to be the information that can be learned from the designated output of the computation.

Almost all techniques rely on secure multi party communication protocol. Secure multi party communication is defined as a computation protocol at the end of which no party involved knows anything else except it's own inputs the results, i.e. the view of each party during the execution can be effectively simulated by the input and output of the party. The work on secure multi party communication demonstrated that a wide class of functions can be computed securely under reasonable assumptions without involving a trusted third party.

Cryptographic research typically considers two types of adversaries: A semi-honest adversary is a party that correctly follows the protocol specification, yet attempts to learn additional information by analysing the messages received during the protocol execution. On the other hand, a malicious adversary may arbitrarily deviate from the protocol specification. For example, consider a step in the protocol where one of the parties is required to choose a random number and broadcast it. If the party is semi-honest then we can assume that this number is indeed random. On the other hand, if the party is malicious, then he might choose the number in a sophisticated way that enables him to gain additional information. It is of course easier to design a solution that is secure against semi-honest adversaries, than it is to design a solution for malicious adversaries. A common approach is therefore to first design a secure protocol for the semi-honest case, and then transform it into a protocol that is secure against malicious adversaries. More efficient transformations are often required, since this generic approach might be rather inefficient and add considerable overhead to each step of the protocol.

The main building block oblivious transfer

Oblivious transfer is a basic protocol that is the main building block of secure computation. Oblivious transfer is sufficient for secure computation in the sense that given an implementation of oblivious transfer, and no other cryptographic primitive, one could construct any secure computation protocol.

The protocol involves two parties, the sender and the receiver. The sender's input is a pair (x0, x1) and the receiver's input is a bit {0, 1}. At the end of the protocol the receiver learns the bit (and nothing else) and the sender learns nothing. It is known how to design oblivious transfer protocols based on public key cryptosystems. In the case of semi-honest adversaries, there exist simple and efficient protocols for oblivious transfer.

Oblivious polynomial evaluation: This is another useful cryptographic tool involving two parties. The sender's input is a polynomial Q of degree k over some finite field F (k is public). The receiver's input is an element $z \in F$. The protocol is such that the receiver learns Q (z) without learning anything else about the polynomial and the sender learns nothing.

The most recent and more effective techniques include the Integer Partitioning Based Encryption.

There has been a wide area of research going on towards privacy preservation of data. This method**, Integer Partitioning Based Encryption (IPBE)** does not use a simple perturbation or straight forward public key encryption. The method groups the data into various classes and the encryption is based on the key values generated within each class. Since the key is not a constant private or public key, the method provides a greater amount of protection compared to usual cryptographic techniques. The usual disadvantages of other privacy protection algorithms are overcome in this methodology.

The method should preserve the privacy of the data by not affecting the quality of data. After applying the algorithm the method should be usable for mining.

The actual dataset is having *three kinds of attributes:*

1) The individual specific attributes like Name, SSN Number, Bank account number, etc., are removed as they are not of use in mining.

2) The next set of attributes called **quasi identifiers** which can be collectively used for identifying the records.

3) The remaining attributes are called **sensitive attributes** which have to be protected for security. The task of privacy preserving algorithms is to protect the identification of sensitive attribute value for the corresponding quasi identifier attributes.

Many of the Privacy preserving algorithms alter the quasi identifiers in order to make them not identify the particular sensitive data. They are perturbation based approaches which uses suppression, generalization and others. They alter the values using statistical measures or small random values or fuzzy values. The methods produce an unreal data which are not reversible. The methods alter the data in such a way that they are not reversible and there is no chance for getting cent percent data quality as the original data. These procedures do not produce fruitful result as the altered data are still vulnerable to attacks and are affecting the quality of data very much.

Therefore, it is crucial to standardize the Privacy-preserving data solutions that use both randomization and cryptography in order to gain the advantages of both. This method uses a generated key using which the sensitive values are hidden. Only to the intended user the process is discussed and the anonymous data is reversed to obtain the original data.

Hence, a methodology called Integer Partitioning was developed, where in a number is divided into its additional facts and using that the data bins are divided into various sized classes. Since the encryption is based on the mean value of a class the size of the class is crucial which in turn depends on the partitioning of the bins.

This situation motivated the requirement of a methodology combining both perturbation and cryptography which takes the advantages of both and disadvantages of any.

The working is as follows:

- For the whole database T , the records are divided into bins, having C records in each bin.
- Each bin is divided into r classes.
- This is done by a random value generator from the list of possible combinations.
- The key value is chosen as the decimal number specifying the division of the bin into r classes. Hence, there will be different key value for each bin.

6 single digit numbers are determined to get a sum of 40. Hence, a large no.of possible combination exists.

The Coopetative Model

In this model [13], Data holders that participate in distributed data mining have naturally an interest in the result of the data mining. They are, however, understandably reluctant to share their private data with others to either protect their interests or meet privacy requirements imposed by authorities and/or clients. Data holders, in other words, are ready to cooperate with each other to extract useful information from combined data while competition among them dictates that individual data is not revealed to others.

The term coopetation is used in economics to refer to cooperation between competing entities to improve the overall value of their market. This is quite similar to the distributed data mining scenario where data holders behave with similar motivations. In the coopetative model data holders provide inputs to a relatively small set of data mining servers or so called third parties, which are assumed semi-honest (i.e. they are honest but curious; they

follow the protocol steps, but are interested in any leaked information). Some of the data holders can actively participate in the distributed data mining playing the role of third parties. The non-collusion property must be satisfied by certain sets of third parties. At the end of the protocol the data miner, which can be either a separate entity or one of the data holders, will have
the outcome as an output.

Some of the benefits of the coopetative model are:

- Very efficient data mining protocols can be constructed.
- The major workload can be put on a small set of dedicated servers which are better protected and regulated.
- Only these small sets of servers need to posses the necessary hardware, software and know-how to perform data mining.

The basic version of the coopetative model requires two dedicated third parties and a miner. Data holders secret shares their data and send each share to one of the third parties. For sake of simplicity we can assume that the private input of each data holder is an integer x and the data holder creates two shares r and x − r where r is a randomly selected integer. The share r is sent to the first third party and the share x−r is sent to the other. Clearly both shares are random when observed alone and no single entity (adversary, third party, or miner) can obtain any information about the private input x. The private input can only be revealed when two shares are put together, which never happens in the coopetative model. The third parties work on the individual shares and compute algebraic operations such as numerical difference and comparison on the shares, which are the fundamental operations in many data mining applications (e.g. constructing decision trees, association rule mining and clustering). The results of these operations are the shares of the final outcome of the computation, which can be obtained only by the data miner.

In order for the third parties to work on shares, they need to employ secret sharing schemes which are homomorphic with respect to the operations they perform. For instance, additive secret sharing described above is homomorphic with respect to addition (and subtraction): adding shares pairwise gives an additive sharing of the sum of the secrets. Therefore, the additive secret sharing scheme can directly be used in numerical difference operations in clustering algorithms.

VI. Applications Of Privacy Preserving Techniques

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope.

Medical Databases: There are systems designed for de-identification of clinical notes and letters which typically occurs in the form of textual data. Clinical notes and letters are typically in the form of text which contain references to patients, family members, addresses, phone numbers or providers. Traditional techniques simply use a global search and replace procedure in order to provide privacy. However clinical notes often contain cryptic references in the form of abbreviations which may only be understood either by other providers or members of the same institution. Therefore traditional methods can identify no more than 30-60% of the identifying information in the data. This system was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format. The multi-dimensional information may include directly identifying information such as the social security number, or indirectly identifying information such as age, sex or zip-code. The system was designed in response to the concern that the process of removing only directly identifying attributes such as social security numbers was not sufficient to guarantee privacy.

Bioterrorism Applications: In typical bioterrorism applications, analysis of medical data for privacy-preserving data mining purposes takes place. Often a biological agent produces symptoms which are similar to other common respiratory diseases such as the cough, cold and the flu. In the absence of prior knowledge of such an attack, health care providers may diagnose a patient affected by an anthrax attack of have symptoms from one of the more common respiratory diseases. The key is to quickly identify a true anthrax attack from a normal outbreak of a common respiratory disease, In many cases, an unusual number of such cases in a given locality may indicate a bio-terrorism attack.

Therefore, in order to identify such attacks it is necessary to track incidences of these common diseases as well. Therefore, the corresponding data would need to be reported to public health agencies.

However, in the event of suspicious activity, it allows a drill-down into the underlying data. This provides more identifiable information in accordance with public health law.

Homeland Security Applications: A number of applications for homeland security are inherently intrusive because of the very nature of surveillance. Some examples of such applications are as follows:

Credential Validation Problem: This involves matching the subject of the credential to the person presenting the credential. Example, theft of social security numbers presents a serious threat to homeland security. In the credential validation approach, an attempt is made to exploit the semantics associated with the social security number to determine whether the person presenting the SSN credential truly owns it

Identity Theft: A related technology is to use a more active approach to avoid identity theft. The identity angel system, crawls through cyberspace, and determines people who are at risk from identity theft. This information can be used to notify appropriate parties. We note that both the above approaches to prevention of identity theft are relatively non-invasive and therefore do not violate privacy.

Web Camera Surveillance: One possible method for surveillance is with the use of publicly available webcams , which can be used to detect unusual activity.

Video-Surveillance: In the context of sharing video-surveillance data, a major threat is the use of facial recognition software, which can match the facial images in videos to the facial images in a driver license database. While a straightforward solution is to completely black out each face, the result is of limited new, since all facial information has been wiped out. A more balanced approach is to use selective downgrading of the facial information, so that it scientifically limits the ability of facial recognition software to reliably identify faces, while maintaining facial details in images. The algorithm is referred to as *k*-anonymous, and the key is to identify faces which are somewhat similar, and then construct new faces which construct combinations of features from these similar faces. Thus, the identity of the underlying individual is anonymized to a certain extent, but the video continues to remain useful. Thus, this approach has the flavour of a *k*-anonymity approach, except that it creates new synthesized data for the application at hand.

The Watch List Problem: The aim is to view transactional data such as store purchases, hospital admissions, airplane manifests, hotel registrations or school attendance records in order to identify or track these entities. This is a difficult problem because the transactional data is private, and the privacy of subjects who do not appear in the watch list need to be protected. Therefore, the transactional behavior of non-suspicious subjects may not be identified or revealed. Furthermore, the problem is even more difficult if the watch list cannot be revealed to the data holders. The second assumption is a result of the fact that members on the watch list may only be suspected entities and should have some level of protection from identification as suspected terrorists to the general public. The watch list problem is currently an open problem.

Genomic Privacy: Recent years have seen tremendous advances in the science of DNA sequencing and forensic analysis with the use of DNA.DNA data is considered extremely sensitive, since it contains almost uniquely identifying information about an individual. As in the case of multi-dimensional data, simple removal of directly identifying data such as social security number is not sufficient to prevent re identification. It has been shown that a software called *CleanGene* can determine the identifiability of DNA entries independent of any other demographic or other identifiable information. The software relies on publicly available medical data and knowledge of particular diseases in order to assign identifications to DNA entries. The identification is done by taking the DNA sequence of an individual and then constructing a genetic profile corresponding to the sex, genetic diseases, the location where the DNA was collected etc.

VII. Comparative Study

Sr no	*Name*	*Method/Technique*	*Advantage*	*Disadvantage*	*Reference*
1	Integer partitioning based encryption	Encryption	The method groups the data into various classes and the encryption is based on the key values generated within each class. Since the key is not a constant private or public key, the method provides a greater amount of protection.	It involves complex mathematical computations.	ACM
2	Additive Perturbation	Randomization	Data is changed by adding noise to the original data. Identification of data directly is not possible.	The method on its own is weak and does not offer complete reliability, hence it is used in combination with other algorithms.	ACM
3	Perturbation by random projection technique	Randomization	The original record values cannot be easily guessed from the distorted data. It is relatively simple, and does not require knowledge of the distribution of other records in the data	The quality of data is disturbed and the procedure is irreversible. Reconstructions leads to the leakage of Privacy, which relates to the possible risks	ACM
4	Oblivious transfer	Cryptography	Separate parties can jointly compute any function of their inputs, without revealing any other information. It hides all information except for the designated output of the function.	There may exit Corrupted parties, who choose their inputs independently of the honest parties' inputs. This property is crucial in a sealed auction	International Journal of advanced research in Computer Science and software engineering

Sr no	***Name***	***Method/Technique***	***Advantage***	***Disadvantage***	***Reference***
5	k-anonymous method	k-anonimity	It reduces the granularity of data representation. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data.	The technique is susceptible to many kinds of attacks specially when background knowledge is available to the attacker.The adversary can use an association between one or more identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field further.	International Journal of application and innovation in Engineering
6	Zero knowledge protocol	Elliptic Curve Cryptography Approach	Allows a proof of the truth of an assertion, while conveying no information about itself other than its actual truth.	A proof is complete only if given an honest prover and an honest verifier, the protocol succeeds with vast probability and is sound if the probability of a dishonest prover is negligible.	International Journal of advanced research in Computer Science and software engineering
7	Coopetative model	Cryptography approach	The major workload can be put on a small set of dedicated servers which are better protected and regulated. Only these small sets of servers need to possess the necessary hardware, software and know-how to perform data mining	The basic version of the coopetative model requires two dedicated third parties and a miner. Data secret holders share their data and send each share to one of the third parties. Hence, breaching of trust may take place.	International Journal of application and innovation in Engineering

VIII. Conclusion

Thus various different and efficient techniques for privacy preservation in incremental data mining are presented. They are argued and compared with various types of methods existing in the privacy domain for the databases or data storehouses. They are helpful for real time preservation of information and have their own suitability areas.

References

[1] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining", Proc. ACM SIGMOD

[2] Agrawal D. Aggarwal C. "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms," ACM PODS Conference.

[3] Benny Pinkas, "Cryptographic Techniques for privacy preserving data mining"

[4] Jishuan Jin, Cris Clifton, "When do mining results violate privacy",ACM

[5] http://www.stat.cmu.edu/~jiashun/Research/Year/KDD04.pdf

[6] http://link.springer.com/chapter/10.1007%2F978-0-387-70992-5_2

[7] http://www.thesij.com/papers/CSEA/2013/July-August/CSEA

[8] http://www.thesij.com/papers/CSEA/2013/July-August/CSEA

[9] http://dl.acm.org

[10] Balamurugan, M.J. Bhuvana,S. Chenthur Pandian, "Privacy Preserved Collaborative Secure Multiparty Data Mining", Journal of Computer Science,2012

[11] Thomas B.Pedersen,Yucel Saygen,Erkay Sava,"Secret Sharing vs Encryption-based Techniques for privacy Preserving Data Mining"

[12] K.malco,Dfonse,"A Method for Preserving Secret Information in the Data Mining through Cryptography"

[13]http://www.ijarcsse.com/docs/papers/Volume_3/8_August2013/V3I8-0242.pdf

[14]http://www.ijarcsse.com/docs/papers/Volume_3/8_August2013/V3I8-0242.pdf Charu C. Aggarwal , "A General Survey of privacy preserving Data Mining models and algorithms.

[15] V.Rajalakshmi, G.S.Anandha Mala, "Integer Partitioning Based Encryption for Privacy Preservation in Data Mining",ACM