DSP Unit 5:

(*). Privacy Preserving Data Mining Techniques:

Introduction:

→ Privacy preserving data mining has become important in recent years because of the increasing ability to store personal data about users

→ Important for the increasing sophistication of data mining algorithms to leverage this information.

→ It has become a prerequisite for exchanging confidential information

→ A number of techniques such as randomization and k-anonymity have been suggested to perform privacy-preserving data mining.

→ The problem has been discussed in several communities like database community, statistical disclosure community and cryptography community.

→ Key directions in the field of PPDM are:

① Privacy Preserving Data Publishing.

② Changing results of Data Mining Applications to preserve privacy.

③ Query Auditing.

④ Cryptographic Method for Distributed Privacy.

⑤ Theoritical Challenges in High Dimensionality.

① Privacy Preserving data Publishing:

→ tend to study different transformation methods associated with privacy.

→ These techniques include Randomization, k-anonymity, l-diversity, etc.

→ A significant challenge is striking a balance between privacy and usefulness of the data.

→ The methods mentioned above are crucial when data needs to be made available for research or analysis while adhering to privacy regulations.

② Changing Data Mining results for Privacy:

→ Data mining is a powerful tool for discovering patterns and insights from data.

→ However, the results of data mining can reveal sensitive information about individuals.
→ To address this, privacy preserving techniques are employed.
→ For example, Association rule hiding method modify the results by hiding certain rules that might compromise privacy.
→ This ensures that these results can be safely shared without revealing sensitive information about individuals.

③ Query Auditing:
→ Auditing involves limiting or adjusting the o/p of DB queries.
→ Can be necessary when sensitive data is involved.
→ Example, a query might return less detailed results to prevent extraction of sensitive information.

④ Cryptographic Methods for Distributed Privacy:
→ In situations where data is distributed across multiple sites, cryptographic protocols are used to ensure privacy.
→ These protocols allow secure computation of functions.

⑤ Theoritical Challenges in High Dimensionality:
→ Real datasets often have a large no. of attributes and this complexity makes it challenging to apply privacy preserving techniques.
→ k-anonymization is not effective with increasing dimensionality since the data can typically be combined with either public or background information to reveal the identity.

(*) Privacy Preserving Data Mining Algorithms:

① Statistical Methods for disclosure control:
→ These methods include k-anonymity, swapping, randomization, micro-

aggregations & synthetic data generation.
→ Designed to give an overview of common themes in privacy-preserving data mining.

② Measures of Anonymity:
→ Anonymity is a critical concept in privacy preservation.
→ Methods such as k-anonymity, l-diversity & t-closeness are designed to prevent identification, though the final goal is to preserve the underlying sensitive information.

③ The k-anonymity method:
→ A vital privacy de-identification method.
→ The motivating factor behind this technique is that in certain datasets, there are attributes that on their own may not identify individuals but when combined with publicly available data, they can be used to identify individuals.
→ Example, removing identifiers may still leave attributes like birthdates and zip codes that could be used to reveal identities.

④ The Randomization Method:
→ involves distorting the data to create private representation of records.
→ In most cases, individual records cannot be recovered but aggregate distributions can be recovered that make data mining possible
→ 2 types of perturbations are used:
(a) Additive Perturbation: When randomized noise is added to data
(b) Multiplicative Perturbation: Perturbs data through random projection or rotation.

⑤ Quantification of Privacy:
→ Involves assessing how much risk is there of someone finding out our private information when we use a certain method.

⑥ Utility based Privacy-Preserving Data Mining:
→ Most PPDM methods often reduce data effectiveness when applied to data mining algorithms.

→ Balancing privacy & data accuracy depends on specific privacy preservation algorithm used.
→ Challenge is to maintain maximum data utility without compromising underlying privacy constraints.

7. Mining Association Rules under Privacy Constraints:
→ Ensure no rules leak sensitive data in O/P results.
→ This problem is known as Association rule hiding by DB community & contingency table privacy-preservation by statistical community.

8. Cryptographic methods for Information sharing and Privacy:
→ Secure cryptographic protocols are used when multiple parties need to share aggregate private data without revealing individual details.
→ 2 ways data is distributed among different sites: Horizontal and Vertical partitioning.

9. Privacy Attacks:
→ Useful to examine different ways in which one can make attacks on privacy-transformed data.
→ Attacks include SVD-based methods, spectral filtering & background knowledge attacks.
→ Understanding these attacks help designing more effective Privacy preservation methods.

10. Query Auditing and Inference control:
→ Open to Query DBs can compromise security as they can use various queries to undermine data privacy.
→ It adds noise to query results.
→ Query auditing denies certain queries to prevent privacy violations.

11. Privacy and Dimensionality Curse:
→ Many methods such as k-Anonymity and randomization are less effective in high dimensional data.
→ HD. poses unique challenges for privacy preservation.

⑫ Personalized Privacy Preservation:

→ Personalized Privacy is necessary to adapt privacy preserving algo accordingly.

⑬ Privacy-Preservation of data Streams:

→ A new challenge in privacy preservation is dealing with data streams that grow rapidly.

→ The problem is quite challenging bc data is being released incrementally & inability to rely on past data history.

(*) Randomization Method:

① Adding Noise: we start with a set of data records called X. Each record in X is considered sensitive because it contains personal info.

→ To protect this data, we add random noise to each record.

→ The noise added is sufficiently large so that the individual record values cannot be determined.

② Noise selection: The noise added to each record is selected from a probability distribution denoted as $f_Y(y)$.

→ Noise components are denoted as $Y_1, Y_2 \ldots Y_N$.

→ New set of distorted records: $x_1 + y_1$, $x_2 + y_2 \ldots x_N + y_N$.

→ This new set is denoted by $z_1, z_2, \ldots z_n$.

→ Each record gets its own unique noise component.

③ Creating Distorted Records:

→ The set of records, Z are distorted and altered due to the added noise.

→ $Z = X + Y$ ⇒ $X = Z - Y$ (It is possible to approximate the original probability distribution X.

④ Privacy Protection: While we cannot retrieve the original data, we can still understand the overall data distribution.

⑤ Quantifying Privacy: To measure how effective this method is at safeguarding privacy, we use a metric that indicates how confidently we can estimate the original attribute values. If we can estimate

the original value with high confidence, it means that the method is effectively preserving privacy.

(*) One key advantage of randomization is that it is relatively simple, and does not require the knowledge of distribution of other records.
→ This is not true for other methods as k-anonymity requires knowledge of other records in the data.
→ Randomization can be implemented at data collection time.
→ Weakness is that it treats all records equally irrespective of their local density.

(*) - Randomization Methods for Data Streams:
→ for scenarios involving data stream where data arrives continuously.
→ Data streams are especially vulnerable to privacy attacks, so additional precautions are needed.

① Multiplicative Perturbations:
→ Instead of just adding noise, we can also use multiplicative perturbations.
→ Based on the concept of multi-dimensional projections to reduce the dimensionality of data.
→ It preserves the inter-record distance.
→ It is not entirely immune to attacks, it can be effective in certain situations.
→ Additive perturbations are safe from attacks.
→ If the attacker has no prior information/knowledge of data, then it is relatively difficult to attack the privacy.
→ With some prior knowledge, 2 kinds of attacks are possible:
(a). Known Input-Output Attack:
In this case, the attacker knows some linearly independent collection of records, and their corresponding perturbed version. In such cases,

linear algebra techniques can be used to reverse engineer the nature of privacy preserving transformation.

(b). Known Sample Attack: In this case, the attacker has a collection of independent data samples from the same distribution from which original data was drawn. In such cases, principal component analysis techniques can be used in order to reconstruct the behaviour of original data.

(2). Data Swapping:

→ Values accross different records are exchanged to protect privacy.

→ A significant advantage of this technique is that it preserves lower-order totals of data, which allows certain computations to be performed accurately without violating privacy.

→ It is often used in combination with other techniques like k-anonymity to enhance privacy protection.

(3). Group Based Anonymization:

→ Randomization is also a weakness because outlier records can often be difficult to mask.

→ Another weakness of randomization is that it does not consider that publicly available records can be used to identify the identity of the owners of the record.

→ A common strategy involves organizing data records into groups and applying privacy protection technique to each group in a way that it is specific to that particular group.

k-Anonymity framework:

→ In many cases, personal identifiers like names and social security numbers are removed from data records. However attributes such as zip code, age, sex can still be used to identify individuals.

→ To reduce the granularity, and achieve k-anonymity, attributes are generalized to reduce identification risk. Example, specific birth dates might be generalized to just the year of birth.
→ Alternatively, attributes can be suppressed, entirely removing them from the data.
→ Such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on transformed data.
→ k-anonymity requires that every record in a dataset is indistinguishable from atleast k other records, providing a level of privacy.
→ It is formally defined ensuring that every combination of values in quasi identifiers (attributes that are not unique identifiers) can be matched to atleast k respondents.
→ Achieving optimal k-anonymity is a challenging problem, proven to be NP hard.
→ But the problem can be solved using a number of heuristic methods.
→ It relies on attribute ordering and discretization to create an index and a set enumeration tree.
→ It can use a number of pruning options for better efficiency.
→ A branch and bound technique can be used to improve the quality of solution during the traversal process through the data.
→ Incognito is another approach for computing k minimal generalizations. It uses bottom-up aggregation along domain generalization hierarchies.
→ Incognito method uses a bottom up breadth-first search technique/ method to generate all possible k-anonymous tables for a given private table.

→ First it checks k-anonymity for all attributes and removes all those generalizations that do not satisfy k-anonymity.
→ Then it computes generalizations in pairs, again pruning all the pairs that do not satisfy k-anonymity constraints.
→ 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.

* Distributed-Privacy Preserving Data Mining :-

→ the problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.

→ the broad approach to cryptographic methods tend to compute functions over inputs provided by multiple recipients without actually sharing the input with each other

→ eg, two party setting Alice and Bob may have 2-inputs x and y and wish to compute function f(x,y) without revealing x or y to each other.
This problem can be generalized across k- ~~parties~~ ~~across~~ parties by designing the k-argument function h(x1 .... xk)

ANKIT
Date
Page

In order to compute the function $f(x, y)$ or $h(x_1 \ldots x_k)$ a protocol have to be designed and the robustness of the protocol depends upon the level of trust one is willing to place on Alice and Bob.

→ this is because protocol may be subjected to various kinds of adversarial behaviour :-

(1) Semi-Honest Adversaries :-

→ does not deviate from protocol and not share any input

(2) Malicious adversaries :-

→ may vary from protocol, may send sophisticated inputs to each other.

* Applications of Privacy Preserving Data Mining :-

1. Medical Database : The Scrub and Datafly System

* Scrub :-

→ The Scrub system was designed to remove personal info. from clinical notes and letters which occurs in the form of textual data.

* Datafly systems :

→ was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format.

(2) Bioterrorism Application :-

→ ~~Biological agent from~~
One patient has Anthrax which has symptoms similar to cough, cold and the flu.
In the absence of prior knowledge of such an attack, health care providers may ~~diagnose~~ diagnose a patient affected by anthrax attack of have symptoms of cough, cold or flue.

→ ∴ The corresponding are need to be reported to public health agency.

(3) Homeland Security Application :-

→ A no. of applications for homeland security are inherently intrusive because of the very nature of surveillance.

→ Examples of such applications :-

(i) Credential validation problem
(ii) Identity theft
(iii) Web camera surveillance.
(iv) Video surveillance.

* The watch list problem :-

→ The government has list of known terrorist or suspected entities which it wishes to attack from population.

→ the aim is to view their personal data to identify or track these entities.
→ this is a difficult problem because these data are private.

(4) Genomic Privacy :-

→ DNA data is considered extremely private or sensitive since it contains almost uniquely identifying info. about an individual.

→ It has been shown that a software called CleanGene can determine the identifiability of DNA entries.

→ Genomic privacy is a concept that concerns the storage, repurposing, provision to 3rd party and display of genetic info.

- ✔ Hybrid Audit
  - Combination of Automatic and Manual Audits

20-01-2020

Dr.B.Muruganantham
AP / CSE /SRMIST

Auditing Classification and Types …

Audit Types

Financial Audit – Ensures that all financial transactions are accounted for an comply with law.

Ex : Companies save all trading transactions for a period of time to comply with government regulations

Security Audit – Evaluates if the system is as secure as it should be. The audit identifies security gaps and vulnerabilities

Ex: Company might ask a hacker to break the company's network system to determine how secure or vulnerable the network is.

Compliance Audit – Verifies that the system complies with industry standards, government regulations, or partner and client policies

Ex: All pharmaceutical companies must keep paper trails of all research activities to comply with industry standards as well government regulations

Auditing Classification and Types …

Operational Audit –Verifies if an operation is working according to the policies of the company

Ex: When a new hire starts work, the HR department provides ID Card, Sign disclosure , Confidentiality papers, tax forms , etc.,

Investigative Audit – Performed in response to an event, request, threat, or incident to verify the integrity of the system.

Ex: Employee might have committed a fraudulent activity

Product Audit – Performed to ensure that the product complies with industry standards. This audit sometimes confused with testing, but it should not be.

A product audit does not include auditing of its functionality but entails how it was produced and who worked on its development.

Preventive Audit – Performed to identify problems before they occur.

Ex: Company should conduct both random and routine audits to verify that the business operations are being performed according to specifications.

Benefits and Side Effects of Auditing

✔ Benefits

- Enforces company policies, government regulations and laws
- Lowers the incidence of security violations
- Identifies the security gaps and vulnerabilities
- Provides an audit trail of activities
- Provides another means to observe and evaluate operations of the audited entity
- Provides the sense or state of security and confidence in the audited entity
- Identifies or removes doubts
- Makes the organisation being audited more accountable
- Develops controls that can be used for purposes other than auditing

The Curse of Dimensionality

- ✔ Many privacy-preserving data-mining methods are inherently limited by the curse of dimensionality in the presence of public information.
- ✔ For example, the technique in analyzes the *k*-anonymity method in the presence of increasing dimensionality.
- ✔ The curse of dimensionality becomes especially important when adversaries may have considerable background information, as a result of which the boundary between pseudo-identifiers and sensitive attributes may become blurred.
- ✔ This is generally true, since adversaries may be familiar with the subject of interest and may have greater information about them than what is publicly available.
- ✔ This is also the motivation for techniques such as *l*-diversity in which background knowledge can be used to make further privacy attacks.

l-diversity:

l-diversity, also written as ℓ-diversity, is a form of group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression. The l-diversity model handles some of the weaknesses in the k-anonymity model. Protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity. The l-diversity model adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism.

t-closeness

t-closeness is a further refinement of l-diversity group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.