PROJECT PROPOSAL AND ABSTRACT

PRIVACY PRESERVING DATA MINING TECHNIQUES

CSE 664

APPLIED CRYPTOGRAPHY AND COMPUTER SECURITY

Submitted by:

SUDHARCHITH SONTY

UBID:50169912

UBMAIL: sudharch@buffalo.edu

PROBLEM STATEMENT:

- The primary task in data mining is the development of models about aggregated data, extracting implicit un-obvious patterns and relationships from a warehouse of data sets. This information can be useful to increase the efficiency of the organization and aids in future planning.
- However, of late, internet phishing caused significant security and economic concerns on the users and enterprises worldwide. Diversified communication channels via internet services such as electronic commerce, online-banking, research, and online trade exploiting both human and software vulnerabilities suffered from tremendous financial loss.
- Some of the concerns that provide the motivation for privacy preserving data mining solutions are listed below:
  - Privacy Concerns
  - Proprietary information disclosure
  - Concerns about Association breaches
  - Misuse of mining
- Preservation of privacy in data mining has emerged as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation, and publishing. Ever-escalating internet phishing posed severe threat on widespread propagation of sensitive information over the web.
- Therefore, enhanced privacy preserving data mining methods are ever-demanding for secured and reliable information exchange over the internet. The dramatic increase of storing customers' personal data led to an enhanced complexity of data mining algorithm with significant impact on the information sharing.
- Amongst several existing algorithms, the Privacy Preserving Data Mining (PPDM) renders excellent results related to inner perception of privacy preservation and data mining.
- So, the question is; Can we develop accurate models without access to precise information in individual data records? And if we do so, some of the aspects that are to be kept in mind are as follows:
  - Restrict Access to data (Protect Individual records).
  - Protect both the data and its source.
- Truly, the privacy must protect all the three mining aspects including association rules, classification, and clustering.

PROPOSED SOLUTIONS:

- Amongst several existing algorithms, the Privacy Preserving Data Mining (PPDM) renders excellent results related to inner perception of privacy preservation and data mining.
- The current privacy preserving data mining techniques are classified based on distortion, association rule, hide association rule, taxonomy, clustering, associative classification, outsourced data mining, distributed, and k-anonymity.
- The approach is to implement some of these techniques using existing Cryptographic encryption and decryption algorithms in conjunction with Data Mining Techniques to achieve Privacy Preserving Data Mining.

INPUT DATA RANDOMIZATION:

- Applied generally to provide estimates for data distributions rather than single point estimates.
- A user can alter the value provided to the aggregator.
- The alteration scheme should be known to the aggregator.
- The aggregator Estimates the overall global distribution of input by removing the randomization from the aggregate data Randomization.
- Assumptions:
    - ♦ Users are willing to divulge some form of their data.
    - ♦ The aggregator is not malicious but may honestly be curious.

- Value Distortion Method:
    - Given a value, the client can report a distorted value where is a random variable drawn from a known distribution. (Gaussian Distribution).
    - Reconstruction problem can be viewed in in the general framework of the "Inverse Problems" that is describing system internal structure from indirect noisy data. Bayesian Estimation is an Effective tool for such settings.

DECISION TREE ALGORITHM ON THE RANDOMIZED DATA:

- A decision tree is a rooted tree containing nodes and edges.
- Each internal node is a test node and corresponds to an attribute; the edges leaving a node correspond to the possible values taken on by that attribute.
- For example, the attribute "Home-Owner" would have two edges leaving it, one for "Yes" and one for "No".
- Finally, the leaves of the tree contain the expected class value for transactions matching the path from the root to that leaf. Given a decision tree, one can predict the class of a new transaction t as follows.
- Let the attribute of a given node v (initially the root) be labelled A, where A obtains possible values a_1,:::; a_m. Then, as described, the m edges leaving v are labelled a_1,:::; am respectively. If the value of A in t equals a_i, we simply go to the son pointed to by a_i.
- We then continue recursively until we reach a leaf. The class found in the leaf is then assigned to the transaction.
- To induce decision Trees using randomized data, we need to modify the way we choose the split point and the way we partition the data.
- We also need to resolve the choices keeping in mind the reconstruction of the original data set.

CLUSTERING BASED PPDM:

- An equally contributed multiparty k-means clustering is applied on vertically partitioned data, wherein each data site contributed k-means clustering evenly.

- As per to the basic concept, data sites collaborated to encrypt k values (each associated to a distance between the centre and point) with a common public key in each step of clustering.

- Then, it securely compared k values and outputted the index of the minimum without displaying the intermediate values. In some setting, this is practical and more efficient than Vaidya–Clifton protocol.

EXPECTED OUTCOMES:

- An implementation of Distributed K-Means with encryption using a common Public Key.
- An implementation of Randomization on a given data set using the Value Distortion Method proposed.
- An implementation of the Decision Tree Algorithm over the said randomized data.

REFERENCES:

- Privacy Preserving Data Mining - Cynthia Dwork and Frank McSherry
  http://web.stanford.edu/group/mmds/slides/mcsherry-mmds.pdf

- Privacy Preserving Data Mining – Yehuda Lindell, Benny Pinkas
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.9334&rep=rep1&type=pdf

- A comprehensive review on privacy preserving data mining by
  Yousra Abdul, Alsahib S. Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque
  https://springerplus.springeropen.com/articles/10.1186/s40064-015-1481-x

- Privacy Preserving Data Mining – Moheeb Rajab
  http://www.cs.jhu.edu/~fabian/courses/CS600.624/slides/privacy-preserving.pdf

- Privacy Preserving Data Mining – Rakesh Agarwal, Ramakrishnan Srikant
  https://www.utdallas.edu/~muratk/courses/privacy08f_files/agrawal00privacypreserving.pdf