RESEARCH ARTICLE OPEN ACCESS

Information Security in Large Amount of Data: Privacy and Data Mining

Prof. Harish Barapatre [1], Ms. Sampada Samel [2]
Department of Computer Science and Engineering
YTIET, Bhivpuri Road
Mumbai - India

ABSTRACT

Now a days, In our day to day life, development in data mining becomes very much popular. But, growing popularity and development in data mining technologies brings serious threat to the security of individual's sensitive information. So, To avoid access to one's sensitive information, Privacy Preserving Data Mining (PPDM) is used. In this technique, data gets modified in order to secure one's sensitive information. PPDM technique is mainly focused on how privacy maintained at Data Mining. However, Sensitive information can be retrieved at Data Collection, Data Publishing and Information Delivering processes. In this Paper, we briefly discuss the Privacy Preserving Data Mining with respect to user such as Data Provider, Data Collector, Data Miner and Decision Make. We will discuss privacy concerns related to each user. We also discuss about the Game Theory approaches.

Keywords:- Data Mining, Privacy, information, KDD, PPDM, etc

I. INTRODUCTION

In recent years, Data mining has gained a lot public-ity. Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data.[1]

1.1. THE PROCESS OF KDD

To retrieve useful information from data, following steps are followed:

Step 1: Data Processing. It includes basic operations like Data selection, Data cleaning and Data integra-tion. Data selection process retrieves the data rele-vant to KDD task from database. Data cleaning pro-cess is used to remove noise and inconsistant data and Data Integration process combines data from multiple sources.

Step 2: Data Transformation. Data Transformation performs feature selection and feature transformation.

Step 3: Data Mining. Data Mining process extracts the useful data from large amount of data.

Step 4: Pattern evaluation and presentation. It in-cludes basic operations like identifying the truly in-teresting patterns which represent knowledge and presenting the mined knowledge in an easy-to-understand fashion.

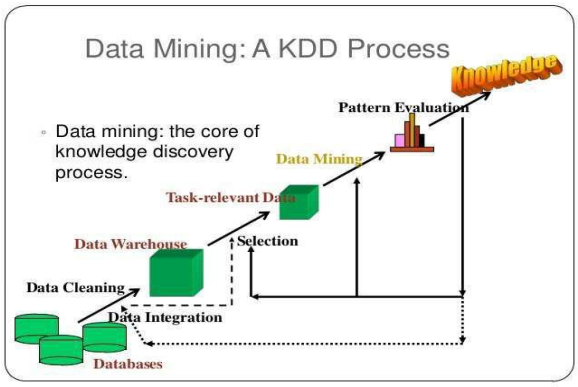


Figure 1. An overview of KDD process

1.2. THE PRIVACY CONCERN AND PPDM

In Recent years, Data Mining becomes very valuable in areas such as Web search, scientific discovery, digital li-braries, etc. But, In Data Mining, individual privacy may be compromised due to unauthourized access to the per-sonal information. Due to unathorized access to one's pri-vate data, people may suffer. So, To avoid this PPDM tech-nique has gained a great development in recent years. The main objective of PPDM (Privacy Preserving Data Min-ing)[2] is to safeguard sensitive information from unnec-essary disclosure of personal information.

1.3. USER ROLE-BASED METHODOLOGY

PPDM technique mainly focus on how to hide sensi-tive information from data mining operations. But, It is also true that security issues arise in data collecting, data processing also. As shown in Fig.1, we can identify differ-ent types of users in data mining scenario.[3]

Data Provider: The user who owns some data that are desired by data mining task.

Data Collector: The user who collects data from data providers and then publish the data to the data miner.

Data Miner: The user who performs data mining tasks on the data.

Decision Maker: The user who makes decisions based on the data mining results in order to achieve certain goals.

We will discuss about each user's privacy concerns one by one.[4]

1.3.1. DATA PROVIDER. Data provider has to ensure that the data provided by him to the data collector must not contain any privacy information. the provider should be able to make his very private data, namely the data con-taining information that he does not want anyone else to know, inaccessible to the data collector. On the other hand, if the provider has to provide some data to the data collec-tor, he wants to hide his sensitive information as much as possible and get enough compensations for the possible loss in privacy.

1.3.2. DATA COLLECTOR. The data received from data provider may contain sensitive information. Directly sending such a data to the data miner, may violate data provider's privacy, hence data modification is required. But major concern of data collector is to guarantee that the modified data contain no sensitive information and still preserve high utility. If data after modification is not useful then collecting data from data provider becomes meaning-less.

1.3.3. DATA MINER. The data miner receives data from data collector and performs different mining algorithms to achieve certain goal. Data miner wishes to extract useful information from the data provider. Data miner has ma-jor responsibility of protecting sensitive data, while data miner can focus on how to hide sensitive information after mining from untrusted parties.

1.3.4. DECISION MAKER. As shown in the fig.2, De-cision maker gets information from Data Miner and from Information transmitter. It is likely that the information transmitter may change mining results intentionally or un-intentionally. So, main concern of Decision maker is that whether mining results are credible.

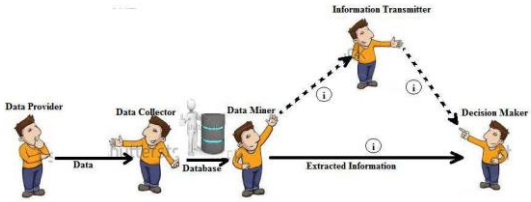


Figure 2. A simple illustration of the application scenario with data mining at the core.

II. DATA PROVIDER

2.1. CONCERNS OF DATA PROVIDER

As shown in fig.2, There are two data providers. Data provider which provides data to the data collector. and Data collector which provides data to the Data miner. We will consider here ordinary Data provider which owns some data which is desired by Data mining process. If the data provider provides his sensitive data to the Data col-lector, then his privacy might get compromised due to un-expected data breach or exposure of sensitive information.

2.2. APPROACHES TO PRIVACY PROTEC-TION

2.2.1. LIMIT THE ACCESS. Data provider can provide his data to the Data collector in active way or passive way.

Active Way: Data provider voluntarily opts in a sur-vey initiated by the Data collector or fill in some reg-istration form to create an account in a website.

Passive Way: Data collector collects the Data provider's data by the Provider's routine activi-ties. Data collector retrieves the data by record-ing provider's routine activities unaware of Data Provider.

Data provider can avoid tracking his routine activities by emptying cache, deleting cookies, clearing usage records of applications, etc. Current security tools that are de-veloped for internet environment to protect provider's data can be categorized into three types:

1. Anti-tracking extensions: Data collector can re-trieve user's sensitive data by tracking his/her routine activities. To avoid this unathourized access to the provider's data, provider can user anti-tracking tools such as Disconnect, Do Not Track Me, Ghostery, etc.
2. Advertisement and script blockers: By adding browser extensions such as AdBlockPlus, NoScript, FlashBlock, etc. user can block advertisements on the sites and kill scripts and widgets that send user's sen-sitive data to unknown third party.

3. Encryption tools: A user can utilise encryption tools such as MailCloak, TorChat to encrypt mails to make sure that a private communication between two par-ties cannot be intercepted by third parties.

In addition to all these tools, user can use anti-virus, anti-malware tools to protect data. Using such a tools user can limit the access of his/her sensitive data to third parties.

2.2.2. TRADE PRIVACY FOR BENEFIT.

In some cases, provider needs to make tradeoff between the loss of privacy and the benefits brought by participating in Data mining. Consider shopping website. If website tracks user's routine activities and find outs user interested prod-ucts, then it will be beneficial for user also. User can fill better shopping experience in this case.Now suppose user has to enter information about salary on shopping website, then the website can show the interested item in user's bud-get. so, disclosure of sensitive information such as salary is more beneficial as it reduces the searching time of the user.

2.2.3. PROVIDE FALSE DATA.

Data providers takes efforts to hide sensitive information from data collector. Data collector takes efforts to hide sensitive information from Data miner. But, In today's internet age, internet users cannot completely stop the unwanted access to user's personal information. So, instead of trying to limit the access, the data provider can provide false information to untrustworthy Data collectors. Following methods can be used to falsify the data.

1. sockpuppets: A sockpuppet is a false online identity. By using multiple sockpuppets, the data produced by one individual's activities will be deemed as data be-longing to different individuals. Data collector do not have enough knowledge to relate different sockpup-pets with one individual.So, user's true activities are unknown to others and user's sensitive information cannot be easily retrieved.
2. Clone identity:This technique can protect user's pri-vacy by creating fake identity.This clone identity au-tomatically makes some actions which are totally dif-ferent from user's actions.So if the third party tries to retrieve user's data, then it will get data from clone identity which is completely different.By this way, user's sensitive data is unaccessible to the unwanted user.
3. MaskMe: By adding MaskMe browser extension user can hide his/her sensitive data. Wherever user perform online transaction, user has to enter his sen-sitive information such as email id, Bank details, etc. Using this extensions, many aliases are created. so, user's sensitive data can be secured.

2.3. CONCLUSION

Data Provider can secure his data by three ways: he can limit the access to his online activities or data by using anti-tracking extensions, advertisement and script block-ers or by using encryption tools to encrypt emails between two private parties.Data Provider can also demand for high price to disclose his private data with others.Nowadays, whatever you try, but the hackers can get your secure information.So, Data provider can provide false data to misguide such a hackers.Using sockpuppet, Data provider can make different sockpuppets. Data provider can use MaskMe to mask his sensitive information.

III. DATA COLLECTOR

3.1. CONCERNS OF DATA COLLECTOR

As shown in Fig.2, Data collector collects the data from Data provider and provide this data to Data Miner. Data collected from Data provider may contain sensitive information of the individual. If such a data is directly send to the Data Miner, then individual's sensitive information disclosed to the unwanted third parties or Data miner. So, before sending data to the Data miner, Data collector has to check whether data contains sensitive information or not.If so then Data collector has to encrypt the data collected from Data provider and then send it to the Data miner.

Data collector has to modify the data before releas-ing it to the Data miner.But, After using modification tech-niques, there will be loss in data utility.So the main con-cern of Data miner is that the data must retained utility after the modification. Otherwise collecting data is waste process.

The data modification process adopted by Data collec-tor with the goal of preserving privacy and utility simul-taneously is called as Privacy Preserving Data Publishing (PPDP)[5].

3.2. APPROACHES TO PRIVACY PROTEC-TION

3.2.1. BASICS OF PPDP. The original data is in the form of table with multiple records.Each record consists of four types of attributes.

1. Identifier (ID): Attributes that uniquely identifies user on cloud
2. Quasi-identifier (QID): Attributes that linked with the external data to re identify user.
3. Sensitive Attribute (SA):Attributes that the user wants to hide for privacy.
4. Non-Sensitive Attribute (NSA): Attributes that user don't matter to disclose with anyother.

Data anonymization is a type of information sanitization whose intent is privacy protection.It is the process of either encrypting or removing personally identifiable information from data sets so that identity and sensitive attribute values hidden from adversaries. Record linkage (RL) refers to the task of finding records in a data set that refer to the

same entity across different data sources (e.g., data files, books, websites, databases). In Attribute linkage (AL), the adversary may not precisely identify the record of the target victim, but could infer his/her sensitive values from the published data, based on the sensitive values associ-ated to the group, that the victim belongs to. In Table linkage (TL), the attack seeks to determine the presence or absence of victim's record in the released table. Prob-abilistic linkage, takes a different approach to the record linkage problem by taking into account a wider range of potential identifiers, computing weighs for each identifier based on it's estimated ability to correctly identify a match or non-match, and using these weighs to calculate proba-bility that two given records refer to the same entity. Dif-ferent privacy models includes k-anonymity, l-diversity, t-closeness, epsilon-differential privacy.

k-anonymity is used for record linkage.

l-diversity is used for preventing record and attribute linkage.

t-closeness is used for preventing attribute and proba-bilistic linkage.

epsilon-differential is used for preventing table and probabilistic linkage.

Among all these, k anonymity is widely used.In k-anonymity, attributes are suppressed or generalized until each row is identical with atleast k-1 other rows. Thus it prevents definite database linkages. K-anonymity guaran-tees that the data released is accurate.

Consider following table which gives idea about k-anonymity. Now consider above table A and table B which denotes Raw table and anonymized table respec-tively. Using K-anonymous technique, Data collector can hide Identifiers and Quasi-identifier fields from third par-ties.As shown in fig.B, quasi-identifier fields such as age, sex zipcode are replaced by either special characters or range values or common attribute. So, by using such anonymous table, adversaries are unable to track particular individual.then, the probability that the individual's record being identified by the adversary will not exceed 1/K.

To satisfy privacy model conditions, following opera-tions can be done.

Generalization: Replace some values in the table with with parent value in the taxonomy of an attribute.

Suppression: Replace some values in the table with a special character ("*"), as shown in the column "Zip-code" in table.B.

Anatomization: De-associates the relationship be-tween two.

Permutation: De-associates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.

Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

(a)

Age	Sex	Zipcode	Disease
[1, 10]	People	1****	HIV
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	bronchitis
[11, 20]	People	2****	HIV
[11, 20]	People	2****	cancer
[11, 20]	People	2****	pneumonia
[11, 20]	People	2****	gastritis
[21, 60]	People	3****	flu
[21, 60]	People	3****	pneumonia

(b)

Figure 3. An example of 2-anonymity where QID = Age,Sex,Zipcode.(a)Original Table (b)2-anonymous table

Perturbation: Replace original data values with some synthetic data values.

But, All these privacy model information results into in-formation loss.

3.2.2. PRIVACY PRESERVING PUBLISHING OF SOCIAL NETWORK DATA. Social network data is al-ways represented in the form of graph.where vertex rep-resents an entity and edge represents the relationship be-tween two entities.So, In case of social network PPDP deals with the anonymized graph data. Anonymizing so-cial network data[6] is much more challenging than that of relational data.

it is much more challenging to model background knowledge of adversaries and attacks about social network data than that about relational data. On re-lational data, it is often assumed that a set of at-tributes serving as a quasiidentifier is used to asso-ciate data from multiple tables, and attacks mainly come from identifying individuals from the quasi-identifier. However, in a social network, many pieces of information can be used to identify individuals, such as labels of vertices and edges, neighborhood graphs, induced subgraphs, and their combinations. It is much more complicated and much more difficult than the relational case.

it is much more challenging to measure the informa-tion loss in anonymizing social network data than that in anonymizing relational data. Typically, the infor-mation loss in an anonymized table can be measured

using the sum of information loss in individual tuples. Given one tuple in the original table and the corre-sponding anonymized tuple in the released table, we can calculate the distance between the two tuples to measure the information loss at the tuple level. How-ever, a social network consists of a set of vertices and a set of edges. It is hard to compare two social net-works by comparing the vertices and edges individ-ually. Two social networks having the same number of vertices and the same number of edges may have very different network-wise properties such as con-nectivity, betweenness, and diameter. Thus, there can be many different ways to assess information loss and anonymization quality

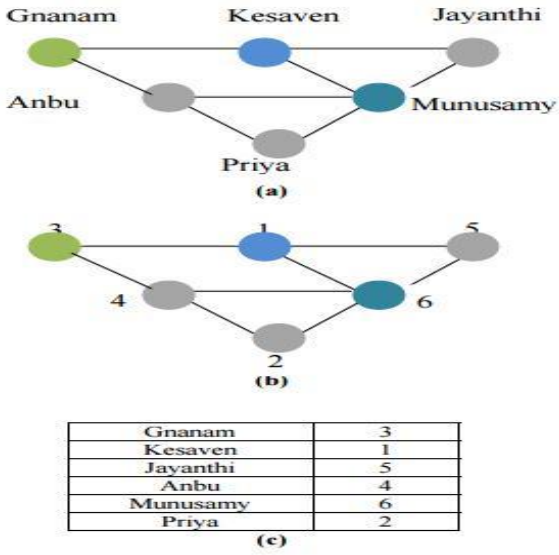
it is much more challenging to devise anonymiza-tion methods for social network data than for rela-tional data. Divide-and-conquer methods are exten-sively applied to anonymization of relational data due to the fact that tuples in a relational tables are separa-ble in anonymization. In other words, anonymizing a group of tuples does not affect other tuples in the ta-ble. However, anonymizing a social network is much more difficult since changing labels of vertices and edges may affect the neighborhoods of other vertices, and removing or adding vertices and edges may af-fect other vertices and edges as well as the properties of the network.

3.2.3. ATTACK MODEL. In Anonymized social net-work data, adversaries often rely on the background knowledge to de-anonymize individuals and learn relation-ships between de-anonymized individuals.

”Seed and Grow” algorithm[7] invented by Peng et al. is used to identify users from an anonymized so-cial graph, based solely on graph structure. The seed stage plants a small specially designed sub graph into undirected graph before its release. After anonymized graph is released , the attacker locates sub graph in anonymized graph.so, the vertices are readily iden-tified and serves as the initial seeds to be grown. The grow stage is essentially comprised of a struc-ture based vertex matching, which further identifies vertices adjacent to initial seeds. This is self reinforc-ing process, in which the seeds grow larger as more vertices are identified.

”Structural Attack”[8] is the attack that de-anonymize social graph data. This attack uses cumulative degree of a vertex.

”Mutual Friend Attack” is de-anonymized data based on the number of social common friends of two di-rectly connected individuals. As shown in Fig.4, The anonymization mapping f, is a random, secret map-ping. Naive anonymization prevents re-identification when adversary has no information about individ-ual in original graph. in practice the adversary may



Gnanam	3
Kesaven	1
Jayanthi	5
Anbu	4
Munusamy	6
Priya	2

(c)

Figure 4. Example of mutual friend attack: (a)Original network; (b)Naive anonymized net-work. (c)Mapping Function (f)

have access to external information about the enti-ties in the graph and their relationships. This infor-mation may be available through a public source be-yond the control of the data owner, or may be ob-tained by the adversarys malicious actions. For ex-ample, for the graph in Figure 1, the adversary might know that Munusamy has three or more neighbors, or that Gnanam is connected to at least two nodes, each with degree 2. Such information allows the adver-sary to reduce the set of candidates in the anonymized graph for each of the targeted individuals. Although an adversary may also have information about the at-tributes of nodes, the focus of this paper is structural re-identification, where the adversarys information is about graph structure. Re-identification with at-tribute knowledge has been wellstudied, as have tech-niques for resisting it. More importantly, many net-work analyses are concerned exclusively with struc-tural properties of the graph; therefore safely publish-ing an unlabeled network is a legitimate goal.

1. PRIVACY MODEL No. of privacy models are pro-posed for graph data based on classic k-anonymity model.[9]

   (a) k-NMF anonymity: It protects the privacy of relationship from the mutual friend attack.

   (b) K^2-degree anonymity: It protects information loss due to friendship attack.

   (c) k-structural diversity anonymization (k-SDA): It protects information loss due to degree attack.

3.2.4. PRIVACY-PRESERVING PUBLISHING OF TRAJECTORY DATA. In recent years, LBS(Location

Based Services)[10] becomes very popular. Using these services user can able to find out interesting places near him/her. If he/she wants information about nearest bank. then he/she can use such a services and able to find out nearest bank location. To provide location-based services, commercial entities (e.g. a telecommunication company) and public entities (e.g. a transportation company) collect large amount of individuals' trajectory data, i.e. sequences of consecutive location readings along with time stamps. If the data collector publish such spatio-temporal data to a third party (e.g. a data-mining company), sensitive infor-mation about individuals may be disclosed. To realize a privacy-preserving publication, anonymization techniques can be applied to the trajectory data set, so that no sensitive location can be linked to a specific individual.

3.3. CONCLUSION

Data Collector receives data from Data provider and sends that data to the Date Miner. Before sending this data to the Data miner, Data collector has to check whether data contains any private information or not. data Collector has to develop different attack models to check whether data contains any private information about data provider

IV. DATA MINER

4.1. CONCERNS OF DATA MINER

Data collector sends the data after modification to the Data miner. Then, the Data miner has to retrieve the im-portant data using different data mining techniques.so,The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To per-form a privacy preserving data mining, the data miner usu-ally needs to modify the data he got from the data collector. As a result, the decline of data utility is inevitable. Similar to data collector, the data miner also faces the privacy util-ity tradeoff problem. But in the context of PPDM, quan-tifications of privacy and utility are closely related to the mining algorithm employed by the data miner.

4.2. APPROACHES TO PRIVACY PROTEC-TION

Privacy preserving data mining approaches are clas-sified into two main categories i.e. Approaches for cen-tralized data mining and Approaches for Distributed data mining.Distributed Data mining again further classified as horizontally partitioned data and vertically partitioned data as shown in Fig.5. Now, most services are using dis-tributed data mining where secure multi-party computa-tion is used. SMS (Secure Multi-party Computation) is a subfield of cryptography.SMC assumes that there are num-ber of participants P_1; P_2; P_3; :::::::; P_m with having private data D_1; D_2; D_3; :::::; D_m respectively.The participants want to compute the value of public function f.We can say that SMC protocal is secure if, at the end of computation, par-

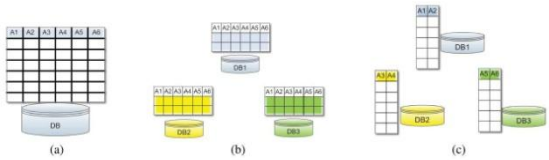


Figure 5. Data Distribution (a)Centralized Data (b)Horizontally Partitioned Data (c)Vertically Partitioned Data.

ticipant can able to view only their own data. So, the main goal of SMC protocol is to find correct data mining results without revealing participants data with others.

4.2.1. PRIVACY-PRESERVING ASSOCIATION RULE MINING. Association rule mining is a two-step process:
(1) Finding all frequent itemsets;
(2) Generating strong association rules from the frequent itemsets.

The purpose of privacy preserving is to discover ac-curate patterns to achieve specific task without precise ac-cess to the original data. The algorithm of association rule mining is to mine the association rule based on the given minimal support and minimal confidence. Therefore, the most direct method to hide association rule is to reduce the support or confidence of the association rule below the minimal support of minimal confidence. With regard to association rule mining, the proposed methodology that is effective at hiding sensitive rules is implemented mainly by depressing the support and confidence.

various kinds of approaches have been proposed to perform association rule hiding.These approaches are clas-sified as five categorize.

Heuristic Distortion Approaches: selects appropri-ate data sets for data modification.

Heuristic blocking approaches: reduces the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data item with specific symbol.

Probabilistic distortion approaches: distorts the data through random numbers generated from prede-fined probability distortion function.

Exact database distortion approaches: formulates the solution of the hiding problem as a constraint sat-isfaction problem (CSP), and apply linear program-ming approaches to it's solution.

Reconstruction-based approaches: generates a database from the scratch that is compatible with a given set of non-sensitive association rules.

4.2.2. PRIVACY-PRESERVING CLASSIFICATION. Data classification is a two step process.

1. Step1 : Learning Step: algorithm generate classifi-cation model.
2. Step2 : Classification: Develops different classifica-tion models such as Decision Tree, Bayesian Model, Support Vector Machine (SMC), etc.

1. DECISION TREE

A decision tree[12] is defined as a predictive mod-eling technique from the field of machine learning and statistics that builds a simple tree-like structure to model the underlying pattern of data. Decision tree is one of the popular methods to classify is able to handle both categorical and numerical data and per-form classification with minimal computation. Deci-sion trees are often easier to understand and compute. Decision tree is a classifier which is a directed tree with a node having no incoming edges called root. All the nodes except root have exactly one incom-ing edge. Each non-leaf node called internal node or splitting node contains a decision and most appropri-ate target value assigned to one class is represented by leaf node. Decision tree classifier is able to di-vide a complex process into number of simpler pro-cesses. The complex decision is subdivided into sim-pler decision on the basis of splitting of complex pro-cess into simple processes. It divides com[plete data set into smaller subsets. Information gain, gain ratio, gini index are three basic splitting criteria to select at-tribute as a splitting point. Decision trees can be built from historical data they are often used for explana-tory analysis as well as a form of supervision learn-ing. The algorithm is designed in such a way that it works on all the data that is available and as per-fect as possible. According to Breiman et al. the tree complexity has a crucial effect on its accuracy perfor-mance. The tree complexity is explicitly controlled by the pruning method employed and the stopping cri-teria used. Usually, the tree complexity is measured by one of the following metrics:

The total number of nodes

Total number of leaves

Tree depth

Number of attributes used

Decision tree induction is closely related to rule in-duction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the an-tecedent part, and taking the leafs class prediction as the class value. The resulting rule set can then be sim-plified to improve its accuracy and comprehensibility to a human user.

2. NAIVE BAYESIAN CLASSIFICATION

The Naive bayesian classifier[13] is a simple but ef-ficient baseline classifier. This classifier used for text

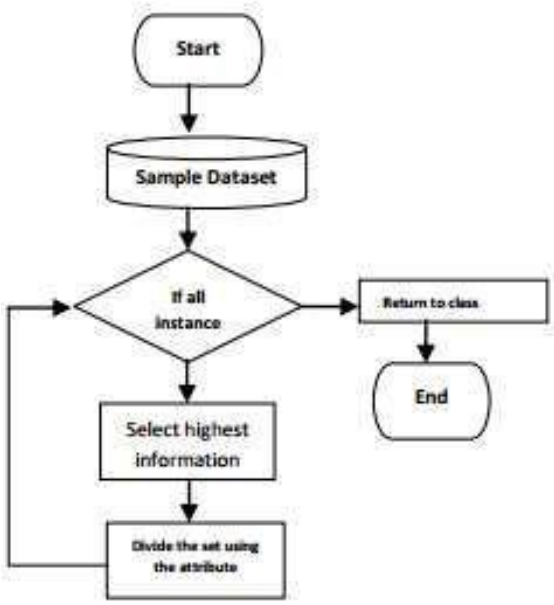


Figure 6. Flowchart for Decision Tree Based Clas-sification.

classification. Naive bayesian is based on a bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. It is simple to compute and computation calculates good results. Thus, preliminary evaluation is carried out using the Naive Bayesian classifier to serve both as a baseline and to decide whether more sophisti-cated solutions are required. The problem of secure distributed classification is an important one. The goal is to have a simple, efficient, easy to compute and privacy-preserving classifier. The ideal would be for all parties to decide on a model. Jointly se-lect/discover the appropriate parameters for the model and then use the model locally as and when necessary. We discuss the specifics in the context of the Naive Bayesian classifier later. in this, data is assumed to be horizontally partitioned. This means that many par-ties collect the same set of information about different entities. Parties want to improve classification accu-racy as much as possible by leveraging other parties data. They do not want to reveal their own instances or the instance to be classified. Thus, what we have is a collaboration for their own advantage. One way to solve this is to decide on a model. The model param-eters are generated jointly from the local data. Clas-sification is performed individually without involving the other parties. Thus, the parties decide on sharing the model, but not the training set nor the instance to be classified. This is quite realistic. For exam-ple, consider banks which decide to leverage all data to identify fraudulent credit card usage, or insurance companies which jointly try to identify high-risk cus-tomers. In this paper, we use / extend several existing

cryptographic techniques to create a privacy preserv-ing Naive Bayesian Classifier for horizontally parti-tioned data

3. SUPPORT VECTOR MACHINE

Support vector Machine (SVM)[14] is one of the most developed classification methodology in data mining. It provides properties such as the margin maximiza-tion and nonlinear classification via kernel tricks and has proven to be effective in many real world applica-tions.Privacy preserving SVM classification solution, PP-SVM which constructs the global SVM classifica-tion model from the data distributed from a multiple parties.The data may be partitioned horizontally, ver-tically or in an arbitry manner between the parties. The data of each party is kept private, while the fi-nal model is constructed at an independent site, This independent site then performs classification of new instances. of sense in many different contexts. For example, consider a clearing house for a consortium of banks. The different banks collect data of their customers. The features collected such as age, gen-der, balance, average monthly income, etc. are the same for all ban k. Thus, the data is horizontally distributed. the clearing house is an independent en-tity, unrelated to any of the banks. The classification model is constructed at the clearing house while pre-serving the privacy of the individual data from each of the banks. When a bank has a new instance it wants to classify, it goes through a secure protocol with the clearing house to classify just this instance. The clear-ing house learns nothing. This would allow all of the banks to leverage the global data without compromis-ing on privacy at all.

4.3. CONCLUSION

Data Miner has to retrieve the important data using different data mining techniques. so, The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To perform a privacy pre-serving data mining, the data miner usually needs to mod-ify the data he got from the data collector. As a result, the decline of data utility is inevitable. Similar to data collec-tor, the data miner also faces the privacy utility tradeoff problem. By using different algorithm techniques such as Decision tree, Support Vector Machine, Naive Bayesian Techniques Data collector modifies data and sends it to the Decision maker.

V. DECISION MAKER

5.1. CONCERNS OF DECISION MAKER

The final goal of data mining process is to provide use-ful information to the decision maker, so that the decision maker can choose a result which is better way to achieve his objective. As we can see, Data provider sends data to Data collector, Data collector sends data to the Data miner and finally Data miner sends data to the Decision Maker. So, we can say that Decision maker is less respon-sible for the data security. The data mining results pro-vided by the data miner are of high importance to the de-cision maker. If the results are disclosed to someone else, e.g. a competing company, the decision maker may suf-fer a loss. That is to say, from the perspective of decision maker, the data mining results are sensitive information. On the other hand, if the decision maker does not get the data mining results directly from the data miner, but from someone else which we called information transmitter, the decision maker should be skeptical about the credibility of the results, in case that the results have been distorted. Therefore, the privacy concerns of the decision maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

5.2. APPROACHES TO PRIVACY PROTEC-TION

5.2.1. DATA PROVENANCE.

Usually, Decision maker receives data from data miner but, in some cases if the decision maker does not get the data mining results di-rectly from the data miner i.e. receives data from other sources ,then he wants to know how the results are deliv-ered to him and what kind of modifications are applied to the results, so that he can decide whether the results are trusted or not. This is why ”provenance” is needed. The term provenance [15] originally refers to the custody of the data. In computer science, data provenance refers to the information that helps determine the derivation history of the data, starting from the original source. Two kinds of information can be found in the provenance of the data: the ancestral data from which current data evolved, and the transformations applied to ancestral data that helped to produce current data. With such information, people can better understand the data and judge the credibility of the data. data provenance has been extensively studied in the fields of databases and work flows. Several surveys are now available. The following five aspects are used to cap-ture the characteristics of a provenance system:

1. Application of provenance. Provenance systems may be applied in many fields to support a number of uses, such as estimate data quality and data relia-bility, trace the audit trail of data, repeat the derivation of data, etc.
2. Subject of provenance. Provenance information can be collected about different sources and at various levels of detail.
3. Representation of provenance. There are mainly two types of methods to represent provenance in-formation, one is annotation and the other is inver-sion. The annotation uses metadata. Using inversion method, derivations are inverted to find out inputs to the derivations.

4. Provenance storage. Provenance is tightly coupled with the data it describes and located in the same data storage system or even be embedded within the data. Alternatively, provenance can be stored sepa-rately with other metadata or simply by itself.
5. Provenance dissemination. A provenance system can use different ways to provide the provenance in-formation, such as providing a derivation graph that users can browse and inspect.

5.2.2. WEB INFORMATION CREDIBILITY.

Be-cause of the lack of publishing barriers, the low cost of dissemination, and the lax control of quality, credibility of web information[16] has become a serious issue. Tudjman et al. identify the following five criteria that can be employed by Internet users to differentiate false information from the truth:

1. Authority: the real author of false information is usu-ally unclear.
2. Accuracy: false information dose not contain accu-rate data or approved facts.
3. Objectivity: false information is often prejudicial.
4. Currency: for false information, the data about its source, time and place of its origin is incomplete, out of date, or missing.
5. Coverage: false information usually contains no ef-fective links to other information online.

5.3. CONCLUSION

the privacy concerns of Decision Maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

VI. GAME THEORY IN DATA PRIVACY

6.1. GAME THEORY PRELIMINARIES

Till now, we have discussed about the privacy con-cerns for each users and control measures taken by each user.Now, we will focus on the iterations among different users.

The main elements of Game theory[17] are players, actions, payoffs and information. Players perform some actions at particular times in the game. As per the per-formed actions, player will receive payoffs. This payoff is depends on the both player's own actions and other player's actions. Information sets represents a player's knowledge about the values of different variables in the game. A player's strategy is a rule that suggests him which action to choose at what time of the game. A strategy pro-file is an ordered set consisting of one strategy for each of the players in the game. An equilibrium is a strategy pro-file consisting of a best strategy for each of the players in the game. The most important equilibrium concept for the majority of games is Nash equilibrium. A strategy profile is a Nash equilibrium if no player has incentive to deviate from his strategy, given that other players do not deviate.

In many fields such as, Computer Science, Eco-nomics, Politics, etc game theory has been successfully applied.

6.2. PRIVATE DATA COLLECTION AND PUB-LICATION

Data collector collects data from Data provider. So, in order to get benefit, Collector may need to negotiate with the Data provider about the "price" of the sensitive data and the level of privacy protection. Adl build a sequen-tial game model to analyze private data collection process. In this model, a end user who wants to buy a data from the data collector, makes a offer to the collector at the be-ginning of the game. If the data collector accepts offer, then he announces some incentives to data provider's data. Before selling the collected data to the end user, the data collector applies anonymization technique to the data and sends anonymized data in order to protect the privacy of data providers at certain level. On the price and incentives offered by the collector, Data provider decides whether to send data or not.

6.3. PRIVACY PRESERVING DISTRIBUTED DATA MINING

6.3.1. SMC-BASED PRIVACY PRESERVING DIS-TRIBUTED DATA MINING

. Secure Multi-party Computation (SMC) is used in pri-vacy preserving data mining[18]. In this computation, a set of distrustful parties, each with a private input, jointly computes a function over their inputs. SMC protocol is established to ensure that each party will get only compu-tation result and their own data will remain unaffected and secure. However, during the execution of the protocol, a party may take one of the following actions in order to get more benefits:

Semi-honest Adversary:one follows the established protocol and follows the rules of that protocol and correctly performs the computation but attempts to analyze others' private inputs to get benefit.

Malicious Adversary:one arbitrarily deviates from the established protocol which leads to the failure of computation so, quality of whole computation result declined.

Collusion:one party colludes with several other par-ties to expose the private input of another party who doesn't participate in the collusion.

Nanvati and Jinwala model the secret sharing in PPDARM as a repeated game, where a Nash equilibrium is achieved when all parties send their shares and attain a non-collusive

behavior. Based on the game model, they develop punish-ment policies which aim at getting the maximum possible participants involved in the game so that they can get max-imum utilities.

6.3.2. RECOMMENDER SYSTEM

. Many online sites asked the user to recommend the prod-uct of that company to the others using FIVE star to re-view that product. The recommendation system predict user's preference by analyzing the item ratings provided by users. So, user can protect his private preference by providing false data i.e. by providing false ratings to the product. however, this may cause decline of the quality of the recommendation. Halkidi et al. employ game theory to address the trade-off between privacy and recommenda-tion. In the proposed game model, users are treated as a players , and the rating data provided to the recommender server are seen as user's strategies. It has been shown that the Nash equilibrium strategy for each user is to declare false rating only for one item, the one that is highly ranked in his private profile and less correlated with items. To find the equilibrium strategy, data exchange between users and the recommender server is modeled as an iterative process. At each iteration, by using the ratings provided by other users at previous iteration, each user computes a rating vector that can maximize the preservation of his privacy, with respect to a constraint of the recommendation qual-ity. Then the user declare this rating vector to the recom-mender server. After several iterations, the process con-verges to a Nash equilibrium.

6.3.3. LINEAR REGRESSION AS A NON-COOPER-ATIVE GAME

. [19]Data collector receives the data from different Data providers. It may be possible that the data provider may add some noise to protect his data, which affects the accu-racy of the model. The interactions among individuals are modeled as a non-cooperative game, where each individ-ual selects the variance level of the noise to minimize his cost. The cost relates to both the privacy loss incurred by the release of data and the accuracy of the estimated lin-ear regression model. It is shown that under appropriate assumptions on privacy and estimation costs.

6.4. DATA ANONYMIZATION

Chakravarthy propose a k-anonymity method which utilizes coalitional game theory to achieve a proper pri-vacy level, given the threshold for information loss. This method consider each tuple in the data table as a player in the game theory, and calculates the payoff to each player according to a concept hierarchy tree (CHT) of quasi-identifiers. The equivalent class in the anonymous table is formed by establishing a coalition among different tuples based on their payoffs. Given the affordable information loss, this method can automatically find the most feasible value of k, while traditional methods need to fix up the value of k before the starting anonymization process.

6.5. ASSUMPTIONS OF THE GAME MODEL

Most of the proposed approaches adopt the following research paradigm.[20]

Define elements of the game such as Players and their actions and payoffs.

Determine weather game is static or dynamic,with complete information or incomplete information.

Solve the game to find equilibrium.

Analyze the equilibriums to find implications.

Apart from these assumptions, we have to make few as-sumptions while developing the game model.But, Too many assumptions will reduce game model's accuracy. also, An improper assumptions may comprise game model's accuracy.

VII. NON-TECHNICAL SOLUTIONS TO PRIVACY PROTECTION

Apart from the technical solutions to privacy protec-tion, there are non-technical solutions. Legislation on pri-vacy protection has been a prime concern with respect to people. Many countries have established laws to avoid such attacks on individual's private information to regu-late acts involving personal information. Despite the many laws and regulations, now a days, the definition of the right to privacy and the boundary of "legitimate" practice on personal data are still vague.

Besides laws and regulation, industry conventions are very important. Agreement between different organiza-tions on how personal data should be received, placed and modified, can help to build a privacy safe environment for data mining applications.Many countries develop different laws and regulations to avoid such a attack on individual's personal information. If someone tries to do so, he/she will have to go through the legal procedure.

VIII. CONCLUSION

In this paper, We discuss about security concerns and privacy preserving techniques of each user such as Data Provider, Data Collector, Data Miner and Decision Maker.

For Data Provider, can secure his data by three ways: he can limit the access to his online activ-ities or data by using anti-tracking extensions, ad-vertisement and script blockers or by using encryp-tion tools to encrypt emails between two private par-ties.Data Provider can also demand for high price to disclose his private data with others.Nowadays, whatever you try, but the hackers can get your se-cure information.So, Data provider can provide false data to misguide such a hackers.Using sockpuppet, Data provider can make different sockpuppets. Data provider can use MaskMe to mask his sensitive infor-mation.

For Data Collector, He receives data from Data provider and sends that data to the Date Miner. Be-fore sending this data to the Data miner, Data col-lector has to check whether data contains any private information or not. data Collector has to develop dif-ferent attack models to check whether data contains any private information about data provider

For Data Miner, He has to retrieve the important data using different data mining techniques. so,The pri-mary concern of data miner is how to prevent sen-sitive information from appearing in the mining re-sults. To perform a privacy preserving data mining, the data miner usually needs to modify the data he got from the data collector. As a result, the decline of data utility is inevitable. Similar to data collec-tor, the data miner also faces the privacy utility trade-off problem.By using different algorithm techniques such as Decision tree, Support Vector Machine, Naive Bayesian Techniques Data collector modifies data and sends it to the Decision maker.

For Decision Maker, the privacy concerns are twofold: how to prevent unwanted disclosure of sen-sitive mining results, and how to evaluate the credi-bility of the received mining results.

REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques (2nd edition).

[2] Rakesh Agrawal and Ramkrishnan Srikant, ”Privacy-Preserving Data Mining” - ’Privacy Preserving Meth-ods’.

[3] Abinaya.B, Dr.Amitha .T, ”Identifying Users from an Anonymous Social Network” in (IJETCSE) Vol-12,Issue-4, Feb-2015.

[4] D.Ganesh, Dr. S.K.Mahendran, ”Security in Data mining using User role based Methodology” [Page 83] in IJETCCT Vol-1, Issue-2, Nov, 2014.

[5] BENJAMIN C. M. FUNG, KE WANG, RUI CHEN and PHILIP S. YU, ”Privacy-Preserving Data Pub-lishing: A Survey of Recent Developments” in ACM Computing Surveys, Vol. 42, No. 4, Article 14, Pub-lication date: June 2010.

[6] Bin Zhou, Jian Pei, WoShun Luk, ”A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data”.

[7] Wei Peng, Feng Li, Xukai Zou and Jie Wu, ”Seed and Grow: An Attack Against Anonymized Social Net-works” - Page 3.

[8] V.Gnanasekar, S.Jayanthi, ”Privacy Preservation of Social Network Data against Structural Attack using K-Auto restructure” - Figure 1: (a) Social Network

[9] (G), (b)Nave Anonymized Network(G),(c)Mapping function (f).

[10] LATANYA SWEENEY, ”k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY” in In-ternational Journal on Uncertainty,Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[11] Giorgos Poulis, Spiros Skiadopoulos, Grigorios Loukides, Aris Gkoulalas-Divanis, ”Select-Organize-Anonymize: A framework for trajectory data anonymization”.

[12] Mohammad Azam Chhipa, Prof. Lalit Gehlod, ”Sur-vey on Association Rule Hiding Approaches” in IJARCSSE Vol-5, Issue-12, Dec, 2015.

[13] Mr. Tanupriya Choudhury, Mrs. Vasudha Vashisht, Prof.Vivek Kumar, Mr. Himanshu Srivastava, ”Data Mining using Decision Tree for Sales Force Opti-mization”, page [517] in IJARCSSE Vol-3,Issue-4, April 2013.

[14] Jaideep Vaidya, Anirban Basu, Basit Shafiq, Yuan Hong, ”Differentially Private Naive Bayes Classifica-tion”.

[15] Robert Burbidge, Bernard Buxton, ”An Introduction to Support Vector Machines for Data Mining”.

[16] Boris Glavic, Javed Siddique, Periklis Andritsos, Re-nee J. Miller, ”provenance for Data Mining”.

[17] Miriam J. Metzger, ”Making Sense of Credibility on the Web: Models for Evaluating Online Informa-tion and Recommendations for Future Research” in

[18] JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 58(13):20782091, 2007.

[19] Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini, ”Privacy Consensus in Anonymization Systems Via Game Theory”.

[20] Jisha Jose Panackal and Dr Anitha S Pillai, ”Privacy Preserving Data Mining: An Extensive Survey” in Proc. of Int. Conf. on Multimedia Processing, Com-munication and Info. Tech., MPCIT.

[21] Stratis Ioannidis and Patrick Loiseau, ”Linear Regres-sion as a Non-Cooperative Game”.

[22] MALVINA TEMA, ”ASSUMPTIONS OF THE GAME MODEL” in International Relations Quar-terly, Vol-5, No.1.(Spring 2014/1).

[23] LEI XU, CHUNXIAO JIANG, JIAN WANG, JIAN YUAN AND YONG REN,”Information Security in Big Data: Privacy and Data Mining”

[24] R K Adl, M Askari, K Barker, and R S-Naini, ”Privacy Consensus in Anonymization Systems Via Game Theory”.