RESEARCH ARTICLE OPEN ACCESS

Privacy of Integrated Data Using Efficient Classification Algorithm

Smitar. Kapoor*, Dr.R.C.Jain**
*(Department of Software Systems, SATI, Vidisha
** (Director, SATI, Vidisha)

ABSTRACT
Data mining is a field where analysis on the basis of certain parameters can be analyzed. During the process of data mining dataset plays an important role on the basis of which analyses can be fulfilled. But during the processing of the dataset security is also important issue, since the chances of data leak or attacks have increased. Hence privacy preservation is important for the security of these datasets. Here in this paper an efficient implementation of privacy preservation of data using horizontal partition id3 classification tree is proposed which can used as an application for the data integration services. Here data can be collected from different sources which can be integrated at the UTP and privacy of this data can be preserved using our proposed algorithm.

Keywords— privacy preservation, integration, PPDM, Anonymity, UTP, classification, decision tree.

I. INTRODUCTION

New dimension of structure Trust (MLT) poses new challenges for perturbation-based PPDM. In distinction to the single-level trust situation wherever just one rattled copy is released, currently multiple otherwise rattled copies of the same knowledge are offered to knowledge miners at completely different sure levels. The more sure an information manual laborer is, the less rattled copy it will access; it's going to even have access to the rattled copies offered at lower trust levels. Moreover, an information manual laborer could access multiple rattled copies through varied alternative means, e.g., accidental escape or colluding with others.

Data perturbation, a widely utilized and accepted Privacy protective data processing (PPDM) approach, tacitly assumes single-level trust on knowledge miners. This approach introduces uncertainty regarding individual values before knowledge is revealed or discharged to 3rd parties for knowledge mining functions [2].

Underneath the only trust level assumption, an information owner generates just one perturbed copy of its knowledge with a hard and fast quantity of uncertainty. This assumption is proscribed in varied applications wherever a data owner trusts the information miners at completely different levels.

Privacy-preserving information publication attracts nice attention of the community in recent years due to the considerations regarding privacy breaching problems in information publication method and to forestall linking attack, a primary attack in information publication, quite a few PPDP strategies are planned, together with Generalization and randomization [3]. Most of them specialize in static past information set publication and can disclose sensitive info once data is re-published.

ATA mining (knowledge discovery from data) is defined as the non-trivial mining of implicit, earlier unknown, and potentially valuable information from large data sets or databases. Advances in hardware technology have increased the capability to store and record personal data about consumers and persons. Personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining helps to achieve data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure.

Privacy is turning into associate in nursing progressively necessary issue in several data processing applications. A malicious knowledge mineworker might have access to otherwise flustered copies of constant knowledge through varied ways, and will mix these various copies to put together and infer further data regarding the first knowledge that the information owner doesn't unleash. This can be referred to as Diversity Attack. A day, users' area unit departs dozens of electronic trails through varied activities like victimization credit cards, swapping security cards, talking over phones and victimization email. In addition, it's a typical that organizations sell the collected knowledge to different organizations, that use this knowledge for his or her own functions.

Organizations are very passionate about data processing in their day activities. Throughout the total of information mining (from assortment of knowledge to discovery of knowledge) these data, which generally contain sensitive individual info like medical and monetary info, typically get exposed to many parties together with collectors, owners, users and miners. Revealing of such sensitive info will cause a breach of individual privacy. Personal info also can be disclosed by linking multiple information bases happiness to massive information warehouses and accessing internet data.

An unwelcome person or malicious knowledge laborer will learn sensitive attribute values like sickness sort (e.g. HIV positive), and financial gain (e.g. AUD 82,000) of a definite individual, through re-identication of the record from AN exposed knowledge set. This has triggered the event of the many privacy protective techniques that attempt to the data patterns while not directly accessing the first knowledge and guarantees that the mining process doesn't get enough information to reconstruct the first data. Knowledge Perturbation could be a common technique in PPDM and perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before knowledge is revealed.

The scope of perturbation-based PPDM is extended to Multi-Level Trust (MLT-PPDM). Even though MLT-PPDM is powerful against diversity attacks, partial info concealment methodologies like random rotation based mostly knowledge perturbation, k-anonymity and retention replacement aren't supported by MLT-PPDM. In addition MLT-PPDM considers solely linear attacks however additional powerful adversaries apply nonlinear techniques to derive original knowledge and recover additional info [4].

The problem of privacy-preserving data processing has become very necessary in recent years as a result of the increasing ability to store personal information concerning users, and also the increasing sophistication info mining algorithms to leverage this information. Variety of techniques like organization and k-anonymity [5] are instructed in recent years so as to perform privacy-preserving data processing. The matter has been mentioned in multiple communities like the info community, the applied math revelation management community and also the cryptography community. In some cases, the various communities have explored parallel lines of labor that are unit quite similar. This theory will try to explore different topics from the perspective of different communities, and will try to give a fused idea of the work in different communities. The key directions in the field of privacy-preserving data mining are as follows:

Privacy-Preserving Data Publishing: These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization, k anonymity and l-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining [6]. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

Varying the outcome of Data Mining Applications to preserve privacy

In several cases, the results of information mining applications like association rule or classification rule mining will compromise the privacy of the info. This has spawned a field of privacy during which the results of information mining algorithms like association rule mining are changed so as to preserve the privacy of the info. A classic example of such techniques is association rule concealment ways, during which a number of the association rules are suppressed so as to preserve privacy.

Query Auditing: Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries. whereas techniques for restricting queries are discussed in [7].

Cryptographic Methods for Distributed Privacy: In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information. A survey of such methods may be found in [8].

Theoretical Challenges in High Dimensionality: Real data sets are usually extremely high dimensional and this makes the process of privacy preservation extremely difficult both from a computational and effectiveness point of view. In [9], the optimal k-anonymization has been shown

PRIVACY-PRESERVING DATA MINING ALGORITHMS

In key stream are mining problems and challenges associated with each problem. The broad topics covered here are :-

Statistical Methods for Disclosure Control.

The topic of privacy-preserving data mining has often been studied extensively by the data mining community without sufficient attention to the work done by the conventional work done by the statistical disclosure control community. This includes methods such as k-anonymity, swapping, randomization, micro aggregation and synthetic data generation. The idea is to give the readers an overview of the common themes in privacy-preserving data mining by different communities.

Measures of Anonymity. There are a very large number of definitions of anonymity in the privacy preserving data mining field. This is partially because of the varying goals of different privacy-preserving data mining algorithms. For example, methods such as k-anonymity, l-diversity and t-closeness are all designed to prevent identification, though the final goal is to preserve the underlying sensitive information. Each of these methods is designed to prevent disclosure of sensitive information in a different way. This compares and contrasts different measures, and discusses the relative advantages of different measures. Thus provides an overview and perspective of the different ways in which privacy could be defined, and what the relative advantages of each method might be.

The k**-anonymity Method.** An important method for privacy de-identification is the method of k-anonymity [10]. 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.

The Randomization Method. The randomization technique uses data distortion methods in order to create private representations of the records [5]. In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Two kinds of perturbation are possible with the randomization method:

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.

Multiplicative Perturbation: In this case, the random projection or random rotation techniques are used in order to perturb the records.

Quantification of Privacy. A key issue in activity of protection of various privacy-preservation ways is that the means within which the underlying privacy is quantified. The thought in privacy quantification is to live the danger of disclosure for a given level of perturbation and its natural exchange with privacy quantification.

Utility Based Privacy-Preserving Data Mining

Most privacy-preserving data mining methods apply a transformation which reduces the effectiveness of the underlying data when it is applied to data mining methods or algorithms. In fact, there is a natural tradeoff between privacy and accuracy, though this tradeoff is affected by the particular algorithm which is used for privacy-preservation. A key issue is to maintain maximum utility of the data without compromising the underlying privacy constraints. The issue of designing utility based algorithms to work effectively with certain kinds of data mining problems is addressed.

Mining Association Rules under Privacy Constraints.

Association rule mining is one in all the vital issues in data processing. There are two aspects to the privacy preserving association rule mining problem: When the input to the data is perturbed, it is a challenging problem to accurately determine the association rules on the perturbed data. A different issue is that of output association rule privacy. In this case, we try to ensure that none of the association rules in the output result in leakage of sensitive data. This problem is referred to as association rule hiding [11] by the database community, and that of contingency table privacy-preservation by the statistical community.

Cryptographic Methods for info Sharing and Privacy.

In several cases, multiple parties might need to share mixture personal knowledge, while not unseaworthy any sensitive info at their finish [7]. As an example, totally different superstores with sensitive sales knowledge might need to coordinate among themselves in knowing mixture trends while not unseaworthy the trends of their individual stores. This needs secure and cryptanalytic protocols for sharing the data across the various parties. The info is also distributed in 2 ways that across totally different sites:

Horizontal Partitioning: during this case, completely different sites might have different sets of records containing identical attributes.

Vertical Partitioning*: during* this case, completely the various sites might have different attributes of similar sets of records. Clearly, the challenges for the horizontal and vertical partitioning case are quite different

Privacy Attacks

It is useful to examine the different ways in which one can make adversarial attacks on privacy-transformed data. This helps in designing more effective privacy transformation methods. Some examples of methods which can be used in order to attack the privacy of the underlying data include SVD-based methods, spectral filtering methods and background knowledge attacks

Query Auditing and Inference Control.

Many personal databases square measure receptive querying, this will compromise the safety of the results when the adversary will use totally different types of queries so as to undermine the safety of the info. For example, a combination of range queries can be used in order to narrow down the possibilities for that record. Therefore, the results over multiple queries can be combined in order to uniquely identify a record, or at least reduce the uncertainty in identifying it.

There are two primary methods for preventing this kind of attack:

Query Output Perturbation

During this case, we have a tendency to add noise to the output of the question lead in order to preserve privacy [12].

Query Auditing: In this case, we choose to deny a subset of the queries, so that the particular combination of queries cannot be used in order to violate the privacy [13].

Privacy and the Dimensionality Curse

In recent years, it has been observed that many privacy-preservation methods such as k-anonymity and randomization are not very effective in the high dimensional case [14].

Personalized Privacy Preservation

In many applications, different subjects have different requirements for privacy. For example, a brokerage customer with a very large account would likely have a much higher level of privacy-protection than a customer with a lower level of privacy protection. In such case, it's necessary to individualize the privacy protection rule. In customized privacy preservation, we have a tendency to construct anonymizations of the information specified totally different records have a unique level of privacy. The method uses condensation approach for personalized anonymization, while the method in [15] uses a more conventional generalization approach for anonymization.

Privacy-Preservation of Data Streams

A new topic within the space of privacy preserving data processing is that of information streams, within which knowledge grows speedily at a limitless rate. In such cases, the matter of privacy-preservation is kind of difficult since the info is being free incrementally. In addition, the fast nature of data streams obviates the possibility of using the past history of the data. We note that both the topics of data streams and privacy-preserving data mining are relatively new, and there has not been much work on combining the two topics. Some work has been done on performing randomization of data streams [16], and other work deals with the issue of condensation based anonymization of data streams.

II. RELATED WORK

In 2012 by Yaping Li et. all assumption and expand the scope of perturbation-based PPDM to construction Trust (MLT-PPDM) and also the additional trusty an information jack is, the less rattled copy of the info it will access. Underneath this setting, a malicious information jack might have access to otherwise rattled copies of constant information through numerous means that, and will mix these numerous copies to conjointly infer extra info regarding the first information that the info owner doesn't shall unleash. Preventing such diversity attacks is that the key challenge of providing MLT-PPDM services. Here address this challenge by properly correlating perturbation across copies at totally different trust levels and prove that our resolution is powerful against diversity attacks with regard to our privacy goal. That is, for information miners World Health Organization has access to an impulsive assortment of the rattled copies, our resolution stops them from conjointly reconstructing the first information additional accurately than the most effective effort exploitation a person copy within the assortment. Our resolution permits an information owner to come up with rattled copies of its data for impulsive trust levels on demand. This feature offers information house owners most flexibility.

Privacy conserving data processing (PPDM) addresses the matter of developing correct models concerning mass knowledge while not access to specific data in individual knowledge record. A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before knowledge area unit printed. Previous solutions of this approach are unit restricted in their inexplicit assumption of single-level trust on knowledge miners and MLT-PPDM permits knowledge homeowners to come up with otherwise discomposed copies of its knowledge for various trust levels. The key challenge lies in preventing the information miners from combining copies at completely different trust levels to collectively reconstruct the initial data a lot of correct than what's allowed by the information owner. [1]. Rakesh Agrawal et. proposed a unique reconstruction procedure to accurately estimate the distribution of original knowledge values. By victimization these reconstructed distributions, we tend to area unit able to build classifiers whose accuracy is appreciated accuracy of classifiers engineered with the initial knowledge. The fundamental premise was that the sensitive values during a user's record are going to be discomposed employing a randomizing operate so they can't be calculable with sufficient exactness. Organizations are often do victimization Gaussian or Uniform perturbations [17].

By Class and Local are both effective in correcting for the effects of perturbation. At 25% and 50% privacy levels, the accuracy numbers are close to those on the original data. Even at 100% privacy, the algorithms were within 5% to 15% (absolute) of the original accuracy. Recall that if privacy were to be measured with 95% confidence, 100% privacy means that the true value cannot be estimated any closer than an interval of width which is the entire range for the corresponding attribute. We believe that a small drop in accuracy is a desirable trade-off for privacy in many situations.

Local performed marginally higher than by Class, however needed significantly a lot of computation. Investigation of what characteristics may create native a winner over by Class (if at all) is associate open downside.

For identical privacy level, Uniform perturbation did significantly worse than mathematician before correcting for organization, however solely slightly worse once correcting for organization. Thus the selection between applying the Uniform or mathematician distributions to preserve privacy ought to be supported different considerations: mathematician provides a lot of privacy at higher confidence thresholds, however Uniform is also easier to elucidate to users [17].

In 2008 BY Benjamin C. M. et. all consistently characterize the correspondence attacks Associate in Nursing propose an economical anonymization algorithmic rule to thwart the attacks within the model of continuous knowledge commercial enterprise. The majority thought of a single static unleash. Such mechanisms solely shield the information up to initially the primary unleash or first recipient. In sensible applications, knowledge is revealed endlessly as new knowledge arrives; a similar knowledge is also anonym zed differently for a different purpose or a special recipient. In such situations, even once all releases square measure properly k anonym zed, the obscurity of a private is also accidentally compromised if recipient cross-examines all the releases received or colludes with alternative recipients. Preventing such attacks, known as correspondence attacks, faces major challenges Associate in nursing formalized notion of attacks and presented a detection methodology and an anonymization algorithmic rule to forestall such attacks. Finally, we tend to show that each the detection and also the anonymization strategies square measure long to manage multiple releases and alternative privacy needs [18].

[19] Presents the primary study to handle each record insertions and deletions in information re-publication. It proposes a replacement privacy notion known as m-invariance: if a record r has been revealed in releases Ri… Rj wherever i < j, then all QID teams containing r should have an equivalent set of sensitive values, known as the signature [19]. This may make sure the intersection of sensitive values over all such teams doesn't cut back the set of sensitive values. To take care of m-invariance, their technique adds counterfeit records appreciate infrequent sensitive values to form those equivalence categories and have an equivalent signature. Counterfeits, however, might not be ac- in some cases. Suppose a drug company needs to investigate patient reaction to bound medication. Infrequent sensitive values like the negative reactions square measure the foremost fascinating ones and therefore the target for analysis. However, with several counterfeit negative reactions that correspond to no real-life entities, it's tough to deploy the results obtained from such information. Note that, even within the “insertion only" case, adding counterfeits remains necessary, for example, once a record with a replacement sensitive price is another. In distinction, our technique guarantees information honesty at record level: every revealed record corresponds to a real-life entity.

In [20] studies the matter of anonym zing consecutive unleashes wherever every future release publishes a different set of attributes for a similar set of records. In distinction, this paper considers every unleash that mixes new records with antecedently collected records over a similar set of attributes. The attack and bar mechanisms are arbitrarily different in these 2 commercial enterprise models. [21] considers the state of affairs that the Case-ID of records should be printed. In our work, we have a tendency to contemplate the state of affairs that the info holder has removed the Case-ID of records, that the attack supported Case-ID [21] doesn't occur. Instead, we have a tendency to affect a brand new variety of attacks notwithstanding no Case-ID is printed. [22] Proposes associate efficient index structure to incrementally k-anonymize every individual unleash, however it doesn't address the correspondence attacks studied during this paper.

In 2012 by M.S. Ramya offers the thought concerning partial info activity methodologies like random rotation perturbation, retention replacement and K obscurity area unit incorporated with MLT-PPDM to reinforce information security and to stop escape of the sensitive information. Finally MLT-PPDM approach is improved to tackle against the non-linear attacks. Privacy protective data processing (PPDM) is employed to extract relevant information from great amount of knowledge and at a similar time defend the sensitive information from the information miners. The matter in privacy sensitive domain is resolved by the event of the Multi Level Trust Privacy protective data processing (MLT-PPDM) wherever multiple otherwise rattled copies of a similar information is accessible to information miners at totally different sure levels. In MLT-PPDM information homeowners generate rattled information by numerous techniques like Parallel generation, sequent generation and On-demand generation and at the tip level Multi-Level Trust in Privacy-Preserving data processing once integrated with partial info activity methodologies facilitate to seek out the correct balance between most analysis results and keep the inferences that disclose non-public info concerning organizations or people at a minimum. Therefore random rotation based mostly information perturbation and K-anonymity area unit incorporated with MLT-PPDM to considerably enhance the information accuracy and to stop the leakage of the sensitive data [4].

III. PROPOSED METHODOLOGY

Input Layer – Input layer comprises of all the parties that are involved in the computation process. They individually calculate the Information Gain of each attribute and send Intermediate result to UTP. This process is done at every stage of decision tree.

Output Layer – The UTP exists at the 2nd layer i.e. the computation layer of our protocol. UTP collects only intermediate results from all parties not data and calculate the total information gain of each attribute. Then find the attribute with highest information gain and then create the root of decision tree with this attribute and send this attribute to all parties for further calculation. This process is also done at every stage of decision tree.

Informal Algorithm

Input Layer

Party individually calculates Expected Information of every attribute.

Party individually calculates Entropy of every attribute.

Party individually calculates Information Gain of each attribute.

Assume there are two classes, *P* and *N*

Let the set of examples *S* contain *p* elements of class *P* and *n* elements of class *N*

The amount of information, needed to decide if an arbitrary example in *S* belongs to *P* or *N* is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Assume that using attribute A set *S* will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$

If S_i contains p_i examples of *P* and n_i examples of *N*, the entropy, or the expected information needed to classify objects in all subtrees S_i is,

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

The encoding information that would be gained by branching on *A*

$$GAIN(A) = I(p, n) - E(A)$$

Output Layer

All party sends Information Gain of each attribute to the UTP

UTP compute the sum of Information Gain of all parties of all attributes (TotalInformationGain( )).

UTP find out the attribute with the largest Information Gain by using MaxInformationGain( )

Create the root with largest Information Gain attribute and edges with their values, then send this attribute to all parties at Input Layer for further development of decision tree.

Recursively do when no attribute is left.

Assumptions

The following assumptions have been set

UTP computes the final result from the intermediate results provided by all parties at every stage of decision tree.

UTP computes attribute with highest information gain and send to all party at every stage of decision tree.

UTP has the ability to announce the final result of the computation publicly.

Each party is not communicating their input data to other party.

The communication networks used by the input parties to communicate with the UTP are secure.

Formal Algorithm

Input Layer

```
Define $P_1$, $P_2$, …., $P_n$ Parties.(Horizontally partitioned).
Each Party contains R set of attributes $A_1$, $A_2$, …., $A_R$.
C the class attributes contains c class values $C_1$, $C_2$, …., $C_c$.
For party $P_i$ where i = 1 to n do
If R is Empty Then
Return a leaf node with class value
Else If all transaction in T($P_i$) have the same class Then
Return a leaf node with the class value
Else
Calculate Expected Information classify the given sample for each party $P_i$ individually.
Calculate Entropy for each attribute ($A_1$, $A_2$, …., $A_R$) of each party $P_i$.
Calculate Information Gain for each attribute ($A_1$, $A_2$,…., $A_R$) of each party $P_i$
Send Information Gain to UTP
End If.
End For
```

Output Layer

```
Computation is done by UTP
 Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
 $A_{BestAttribute}$ ← MaxInformationGain( )
 Let $V_1$, $V_2$, …., $V_m$ be the value of attributes. $A_{BestAttribute}$ partitioned $P_1$, $P_2$,…., $P_n$ parties into m parties
   $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
   $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
   $P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
```

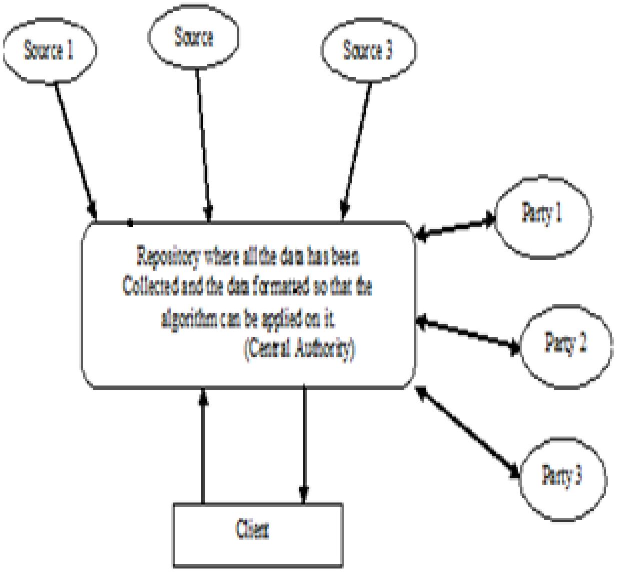
```
Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled $V_1$, $V_2$, …., $V_m$. Such that for every i the edge Vi goes to the Tree
NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
End.
```

1) Algorithm 3 : TotalInformationGain( ) - To compute the Total Information Gain for every attribute.

```
• For j = 1 to R do {Attribute $A_1$, $A_2$,…., $A_R$ }
• Total_Info_Gain($A_j$) = 0
• For i = 1 to n do {Parties $P_1$, $P_2$,…., $P_n$ }
• Total_Info_Gain($A_j$) = Total_Info_Gain($A_j$) + Info_Gain($A_{ij}$)
• End For
• End For
• End.
```

2) Algorithm 4 : MaxInformationGain( ) – To compute the highest Information Gain for horizontally partitioned data.

```
MaxInfoGain = -1
For j = 1 to R do {Attribute $A_1$, $A_2$,…., $A_R$ }
Gain = TotalInformationGain($A_j$)
If MaxInfoGain < Gain then
MaxInfoGain = Gain
$A_{BestAttribute}$ = $A_j$
End If
Return ($A_{BestAttribute}$ )
End For
End.
```


Source 1
Source
Source 3
Repository where all the data has been Collected and the data formated so that the algorithm can be applied on it.
(Central Authority)
Party 1
Party 2
Party 3
Client

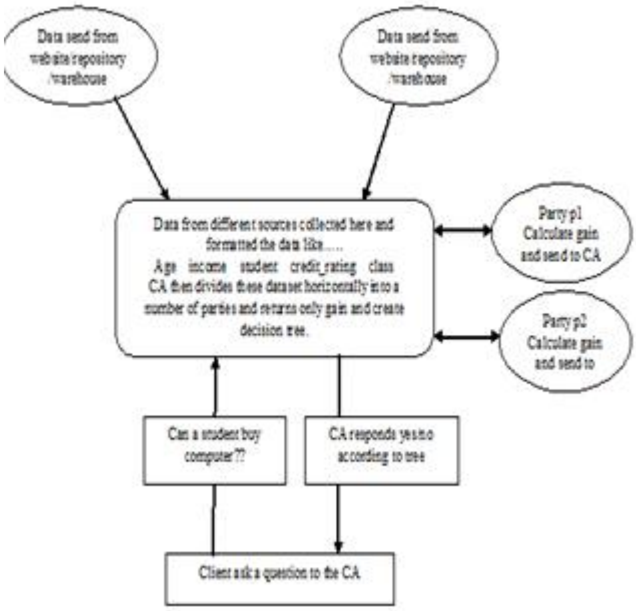
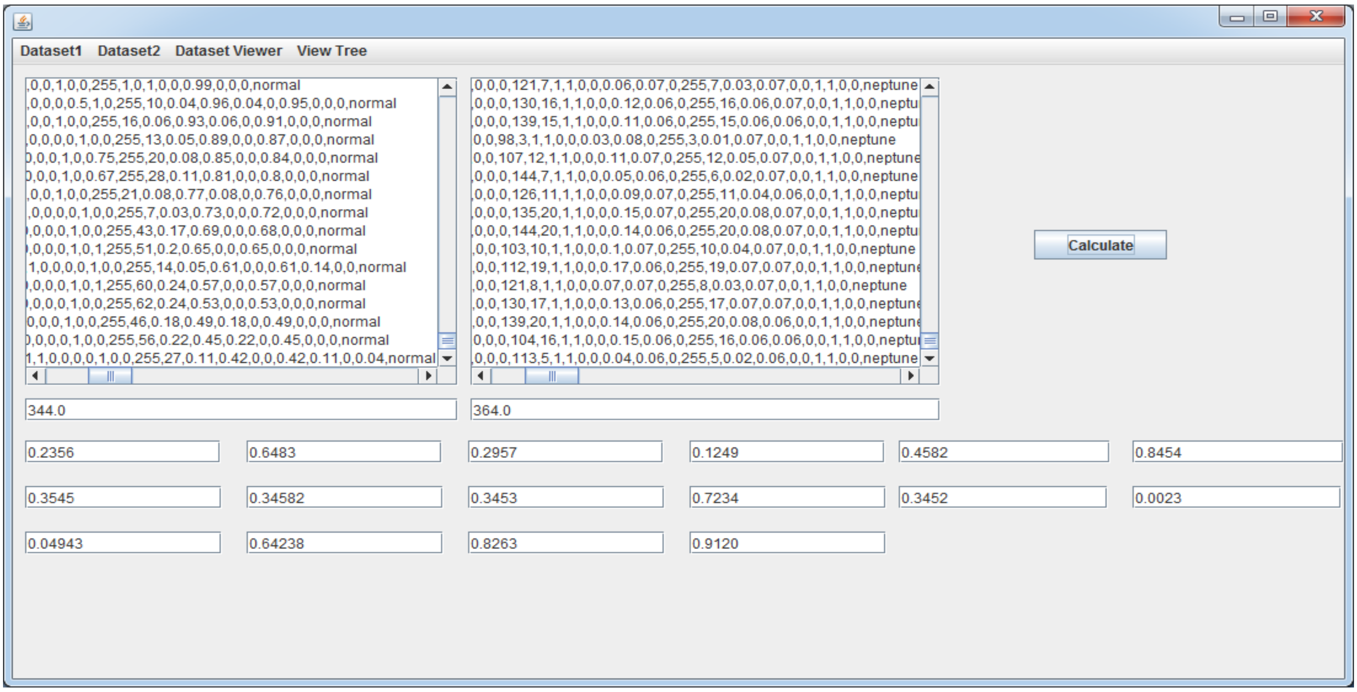

Data send from website repository /warehouse
Data send from website repository /warehouse
Data from different sources collected here and formatted the data like…..
Age income student credit_rating class
CA then divides these dataset horizontally into a number of parties and returns only gain and create decision tree.
Party p1 Calculate gain and send to CA
Party p2 Calculate gain and send to
Can a student buy computer??
CA responds yes/no according to tree
Client ask a question to the CA

Figure 1. Outline of the proposed methodology

Let us take an example of a report of a student where on the basis of the data it can analyze that the particular student can buy a computer or not.

As shown in the fig.2 above is the example working of our proposed methodology, here we use the application of horizontal partition decision tree algorithm for the privacy preservation. The data send from different sources such as websites or any data repository to the UTP where it is collected and these data can be analyzed using our methodology.

RESULT ANALYSIS



```
,0,0,1,0,0,255,1,0,1,0,0,0.99,0,0,0,normal
,0,0,0,0.5,1,0,255,10,0.04,0.96,0.04,0,0.95,0,0,0,normal
,0,0,1,0,0,255,16,0.06,0.93,0.06,0,0.91,0,0,0,normal
,0,0,0,0,1,0,0,255,13,0.05,0.89,0,0,0.87,0,0,0,normal
0,0,0,1,0,0.75,255,20,0.08,0.85,0,0,0.84,0,0,0,normal
0,0,0,1,0,0.67,255,28,0.11,0.81,0,0,0.8,0,0,0,normal
,0,0,1,0,0,255,21,0.08,0.77,0.08,0,0.76,0,0,0,normal
,0,0,0,0,1,0,0,255,7,0.03,0.73,0,0,0.72,0,0,0,normal
0,0,0,0,1,0,0,255,43,0.17,0.69,0,0,0.68,0,0,0,normal
0,0,0,0,1,0,1,255,51,0.2,0.65,0,0,0.65,0,0,0,normal
1,0,0,0,0,1,0,0,255,14,0.05,0.61,0,0,0.61,0.14,0,0,normal
0,0,0,0,1,0,1,255,60,0.24,0.57,0,0,0.57,0,0,0,normal
0,0,0,0,1,0,0,255,62,0.24,0.53,0,0,0.53,0,0,0,normal
0,0,0,1,0,0,255,46,0.18,0.49,0.18,0,0.49,0,0,0,normal
0,0,0,0,1,0,0,255,56,0.22,0.45,0.22,0,0.45,0,0,0,normal
1,1,0,0,0,0,1,0,0,255,27,0.11,0.42,0,0,0.42,0.11,0,0.04,normal
```

```
,0,0,0,121,7,1,1,0,0,0.06,0.07,0,255,7,0.03,0.07,0,0,1,1,0,0,neptune
,0,0,0,130,16,1,1,0,0,0.12,0.06,0,255,16,0.06,0.07,0,0,1,1,0,0,neptu
,0,0,0,139,15,1,1,0,0,0.11,0.06,0,255,15,0.06,0.06,0,0,1,1,0,0,neptu
0,0,98,3,1,1,0,0,0.03,0.08,0,255,3,0.01,0.07,0,0,1,1,0,0,neptune
0,0,107,12,1,1,0,0,0.11,0.07,0,255,12,0.05,0.07,0,0,1,1,0,0,neptune
,0,0,0,144,7,1,1,0,0,0.05,0.06,0,255,6,0.02,0.07,0,0,1,1,0,0,neptune
,0,0,0,126,11,1,1,0,0,0.09,0.07,0,255,11,0.04,0.06,0,0,1,1,0,0,neptu
,0,0,0,135,20,1,1,0,0,0.15,0.07,0,255,20,0.08,0.07,0,0,1,1,0,0,neptu
,0,0,0,144,20,1,1,0,0,0.14,0.06,0,255,20,0.08,0.07,0,0,1,1,0,0,neptu
,0,0,103,10,1,1,0,0,0.1,0.07,0,255,10,0.04,0.07,0,0,1,1,0,0,neptune
,0,0,112,19,1,1,0,0,0.17,0.06,0,255,19,0.07,0.07,0,0,1,1,0,0,neptune
,0,0,121,8,1,1,0,0,0.07,0.07,0,255,8,0.03,0.07,0,0,1,1,0,0,neptune
,0,0,130,17,1,1,0,0,0.13,0.06,0,255,17,0.07,0.07,0,0,1,1,0,0,neptune
,0,0,139,20,1,1,0,0,0.14,0.06,0,255,20,0.08,0.06,0,0,1,1,0,0,neptune
0,0,0,104,16,1,1,0,0,0.15,0.06,0,255,16,0.06,0.06,0,0,1,1,0,0,neptu
,0,0,0,113,5,1,1,0,0,0.04,0.06,0,255,5,0.02,0.06,0,0,1,1,0,0,neptune
```

```
344.0                      364.0

0.2356    0.6483    0.2957    0.1249    0.4582    0.8454
0.3545    0.34582   0.3453    0.7234    0.3452    0.0023
0.04943   0.64238   0.8263    0.9120
```

Figure 3. Result of the privacy preservation

The figure shown above is the selection of dataset and then divided horizontally into two parties and then we choose those two parties to shown attribute values and the different values of the dataset. In this the gain of each attribute is calculated and the attribute having highest gain is the root node of the tree and then second time the gain of each attribute is calculated and the full decision tree is computed.

number_of_data values from source	**Computational Time id3**	**Computational Time Proposed**
24	88	24
40	103	29
50	120	32
150	135	38
250	140	43

Table 1. Comparison of Time Complexity

As shown in the Table is the comparative analysis of time complexity of the existing id3 based decision tree and the horizontal portioning based decision tree. It was found that our proposed algorithm takes very much less time in making of a tree.

number_of_data values from source	**Error rate existing work**	**Error rate Proposed**
24	0.2367	0.1254
40	0.26	0.137
50	0.28	0.165
150	0.24	0.143
250	0.238	0.145

Table 2. Comparison of Error rate

As shown in the Table is the comparative analysis of the mean absolute error of the existing id3 based decision tree and the horizontal portioned base decision tree. Although the difference between the existing and the proposed algorithm is less, but having more absolute error will reduce the efficiency of the algorithm.

IV. CONCLUSION

The horizontal partition based decision tree provides an efficient way of access the data and makes a decision so that the computational time can be reduced. The decision tree made here using the data integrated from the data ware housing should be made secure so that the data when send to the UTP can't be access from the external user. The existing decision tree when applied on the data ware housing takes more computational time and error rate and contains more relative percentage error.

The horizontal partition based decision tree takes less computational cost and the security of the data coming the ware housing takes less error rate. The data integration and sharing across data warehousing provides the security of the data when stored in multiple UTP's. The main conclusion is to integrate the data coming from different UTP's through data warehousing should be secure and create a decision tree.

REFERENCES

[1]. Yaping Li, Minghua Chen, Qiwei Li, And Wei Zhang "Enabling Multilevel Trust In Privacy Preserving Data Mining" , Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, September 2012.

[2] S. Papadimitriou, F. Li, G. Kollios, And P.S. Yu, "Time Series Compressibility And Privacy," Proc. 33rd Int'l Conf. Very Large Data Bases (Vldb '07), 2007.

[3]. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE'05*.

[4]. M.S. Ramya "Partial Information Hiding in Multi-Level Trust Privacy Preserving Datamining" Bonfring International Journal of Software Engineering and Soft Computing, Vol. 2, Special Issue 1, February 2012.

[5] Agrawal R., Bayardo R., Faloutsos C., Kiernan J., Rantzau R., Srikant R.Auditing Compliance via a hippocratic database. VLDB Conference, 2004.

[6] Aggarwal G., Feder T., Kenthapadi K., Khuller S.,Motwani R., Panigrahy R., Thomas D., Zhu A.: Achieving Anonymity via Clustering. ACM PODS Conference, 2006.

[7] AggarwalG., Feder T., KenthapadiK.,MotwaniR., PanigrahyR., Thomas D., Zhu A.: "Anonymizing Tables", ICDT Conference, 2005.

[8] AggarwalG., Feder T.,Kentha padiK., Motwani R., Panigrahy R., Thoma D.,ZhuA.: Approximation Algorithms for k-anonymity. Journal ofPrivacy Technology, paper 20051120001, 2005.

[9] Aggarwal C. C., Yu P. S." On Anonymization of String Data. SIAM Conference on Data Mining, 2007.

[10]. Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules, Workshop on Knowledge and Data Engineering Exchange, 1999.

[11] Bawa M., Bayardo R. J., Agrawal R.: Privacy-Preserving Indexing of Documents on the Network. VLDB Conference, 2003.

[12] Aggarwal C. C., Yu P. S.: A Condensation approach to privacy preserving data mining. EDBT Conference, 2004.

[13] Aggarwal C. C., Yu P. S.: On Variable Constraints in Privacy-Preserving Data Mining. *SIAM Conference,* 2005

[14] AggarwalC., Pei J.,ZhangB.AFramework for Privacy Preservation against Adversarial Data Mining. ACM KDD Conference, 2006.

[15] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k- Anonymization. Proceedings of the ICDE Conference, pp. 217–228, 2005.

[16] Aggarwal C. C.: On Randomization, Public Information and the Curse of Dimensionality. ICDE Conference, 2007.

[17]. Rakesh Agrawal, Ramakrishnan Srikant" Privacy-Preserving Data Mining" IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120, 1999.

[18]. Benjamin C. M. Fung Ke Wang*y* Ada Wai-Chee Fu*x* Jian Pei*y* "Anonymity for Continuous Data Publishing*" EDBT'08,* March 25–30, 2008, Nantes, France. ACM 978-1-59593-926. 2008

[19] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, June 2007.

[20]. K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *SIGKDD*, pages 414{423, August 2006.

[21] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In*SSDBM*, Ban®, Canada, 2007.

[22] T. Iwuchukwu, D. J. DeWitt, A. Doan, and J. F. Naughton. K-anonymization as spatial indexing: Toward scalable and incremental anonymization. In *ICDE*, 2007.