INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

# MULTI PARTY PRIVACY PRESERVING DECISION TREE FOR HORIZONTALLY PARTITIONED DATA

Shaikh Imtiyaj[1*]
[1*]C V Raman College of Engineering, BPUT, Bhubaneswar, Odisha, India
Correspondence Author: mfeedu@gmail.com

**Keywords:** Data mining, Privacy, Data partitioning, ID3 Algorithm, Horizontally partitioned

## Abstract

Data mining is the extraction of hidden predictive information from large databases and also a powerful new technology with great potential to analyze important information in their data warehouses. In this research work, we discussed methods for distributed privacy-preserving mining, and the methods for handling horizontally partitioned data. The primary contribution of this work is to propose a multi party privacy preserving decision tree for horizontally partitioned data by using ID3 algorithm. This has particular relevance to privacy-sensitive searches, particularly top-k queries, and meshes well with privacy policies. There remain many open problems in developing secure solutions based on efficient non secure query processing algorithms.

## Introduction

The networking and databases technologies enable data to be distributed across multi parties and gathered for sharing information. With the rapid growth of the Internet, there is much need to cooperate mining data on the joint databases of multi-participants. Data mining is the extraction of hidden predictive information from large databases and also a powerful new technology with great potential to analyze important information in their data warehouses. Privacy preserving data mining is a latest research area in the field of data mining which generally deals with the side effects of the data mining techniques. Privacy is defined as "protecting individual's information". Protection of privacy has become an important issue in data mining research. Sensitive protection is novel research in the data mining research field.

Privacy Preserving Data mining techniques depends on privacy, which captures what information is sensitive in the original data and should therefore be protected from either direct or indirect disclosure. Secrecy and anonymity are useful ways of thinking about privacy. This privacy should be measureable and entity to be considered private should be valuable.Distributed data mining such as association rule mining and decision tree learning are widely used by global enterprises. Data mining generally assumes a centralized server that collects data from multiple parties before performing data mining on the server. It generally assumes that data on the server can be shared among several parties. Privacy-preserving data mining (PPDM) was introduced to enable conventional data mining techniques to preserve data privacy during the mining process.

Privacy preserving data mining is a new research direction in data mining and knowledge discovery. The main reason for the rapid development of this research area is the growing awareness of the accumulation of huge amounts of easily available data on the Internet – data that may involve a threat to the privacy of users.

Privacy Preserving is the relationship between collection and dissemination of data, technology, the public expectation of privacy, and the legal and political issues surrounding them. Privacy concerns exist wherever personally identifiable information is collected and stored in digital form or otherwise. Improper or non-existent disclosure control can be the root cause for privacy issues. Privacy issues can arise in response to information from a wide range of sources, such as: Healthcare records, Criminal justice investigations and proceedings, financial institutions and transactions, Biological traits, such as genetic material, Residence and geographic records .The challenge in privacy preserving is to share data while protecting personally identifiable information. The fields of data security and information security design and utilize software, hardware and human resources to address this issue. The main goal of Privacy preserving is to mine the rules or pattern accurately without revealing any other private information.

## Data mining

Data mining (knowledge discovery from data) is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Data mining is also known as Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

**Knowledge discovery (KDD) process**

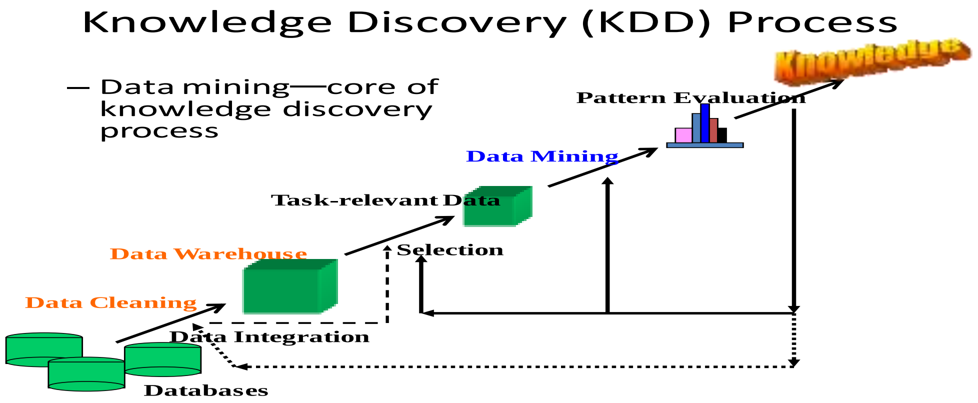


*Figure 1.1: KDD Process*

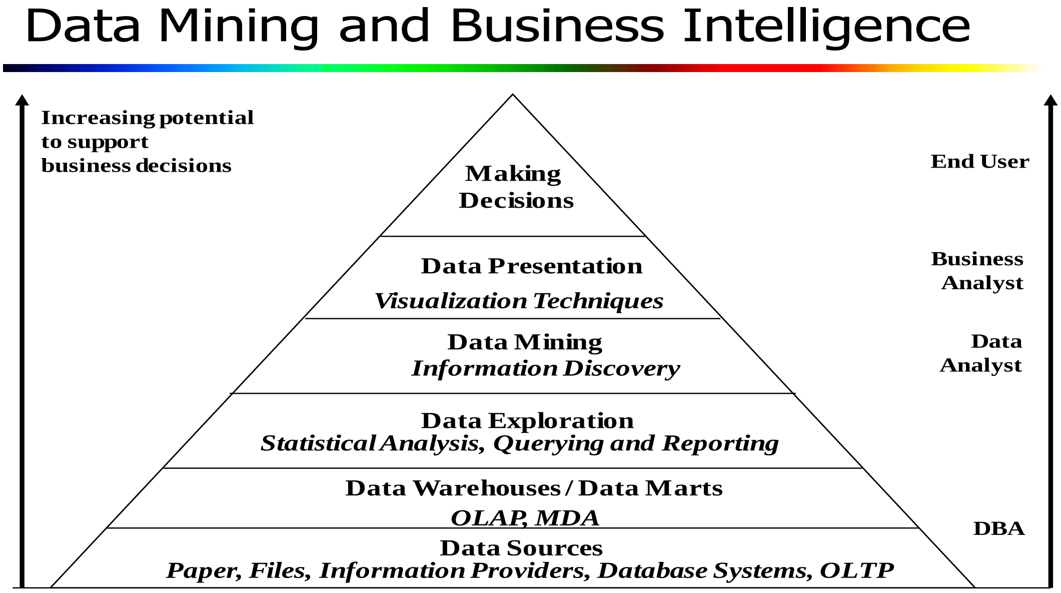

Data Mining and Business Intelligence
Increasing potential to support business decisions
Making Decisions
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Analysis, Querying and Reporting
Data Warehouses / Data Marts
OLAP, MDA
Data Sources
Paper, Files, Information Providers, Database Systems, OLTP
End User
Business Analyst
Data Analyst
DBA

*Figure 1.2: DM and Business Intelligence*

## Features

- Secure communication
- Better quality
- Distributed
- Decentralized and hence robust
- Cost Effectiveness
- High Performance Rate
- Reliability
- High Utilization and Efficiency

## Literature survey

A survey on Privacy-preserving data mining finds numerous applications in surveillance which is naturally supposed to be "privacy-violating" applications. The key is to design methods [1,2] which continue to be effective, without compromising security, a number of techniques have been discussed for bio-surveillance, facial de-dentification, and identity theft. More detailed discussions on some of these issues may be found in [3,4]. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Data mining generally assumes data on the server can be shared among several parties and as privacy issues become more prevalent. Privacy-preserving data mining was introduced [5], [6][7][8] to enable conventional data mining techniques to preserve data privacy during the mining process. Some work has been done to explore privacy-preserving data mining on horizontally and/or vertically partitioned data involving multiple parties so that no single party holds the overall data [9][10][11]. In horizontally partitioned data two or more parties hold different objects for the same set of attributes. It means each object in the virtual database is completely owned by one party. For vertically partitioned data, two parties or more hold the different set of attributes for the same set of objects. In arbitrarily partitioned data, different disjoint portions are held by different parties. This is perhaps the most general form of data partitioning, as introduced by Jagannathan and Wright [12] for two parties. As argued by the authors, although extremely "patch worked" data is unlikely in practice, it is better suited to practical settings as a more general model of horizontally and vertically partitioned data. The secure scalar product is a core operation in decision tree induction for vertically partitioned data. Much work has been done that discussed how the secure scalar product can be computed for two parties .Vaidya and Clifton introduced the Secure Set Intersection Cardinality method to perform secure scalar product for multiple parties [13]. This method has been applied to perform decision tree induction [14] and association rule mining for vertically partitioned data, and SVM model construction for horizontally partitioned data. A major weakness of the Secure Set Intersection Cardinality is its computational and communication complexities, which are O(mn) and O(mn2) respectively, where n is the number of parties and m is the length of private vectors. In arbitrary partitioned the decision tree induction can be performed data partition involving two parties, which is similar to the case for vertically partitioned data. Then, extended to n parties and propose a protocol to securely compute PSP with computational and communication complexities of O(n) and O(mn) respectively.

In horizontally partitioned data sets, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. Many of these techniques use specialized versions of the general methods discussed in [15, 16] for various problems. The work in [17] discusses the construction of a popular decision tree induction method called ID3 with the use of approximations of the best splitting attributes. Subsequently, a variety of classifiers have been generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier [18], and the SVM Classifier with nonlinear kernels [19]. An extreme solution for the horizontally partitioned case is discussed in [20], in which privacy preserving classification is performed in a fully distributed setting, where each customer has private access to only their own record. A host of other data mining applications have been generalized to the problem of horizontally partitioned data sets. These include the applications of association rule mining, clustering and collaborative filtering. A related problem is that of information retrieval and document indexing in a network of content providers. This problem arises in the context of multiple providers which may need to cooperate with one another in sharing their content, but may essentially be business competitors. it has been discussed how an adversary may use the output of search engines and content providers in order to reconstruct the documents. Therefore, the level of trust required grows with the number of content providers. A solution to this problem constructs a centralized privacy-preserving index in conjunction with a distributed access control mechanism. The privacy-preserving index maintains strong privacy guarantees even in the face of colluding adversaries, and even if the entire index is made public.

## Privacy preserving data mining models and methods

### Architecture of data mining system

The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. The architecture of data mining is shown below:

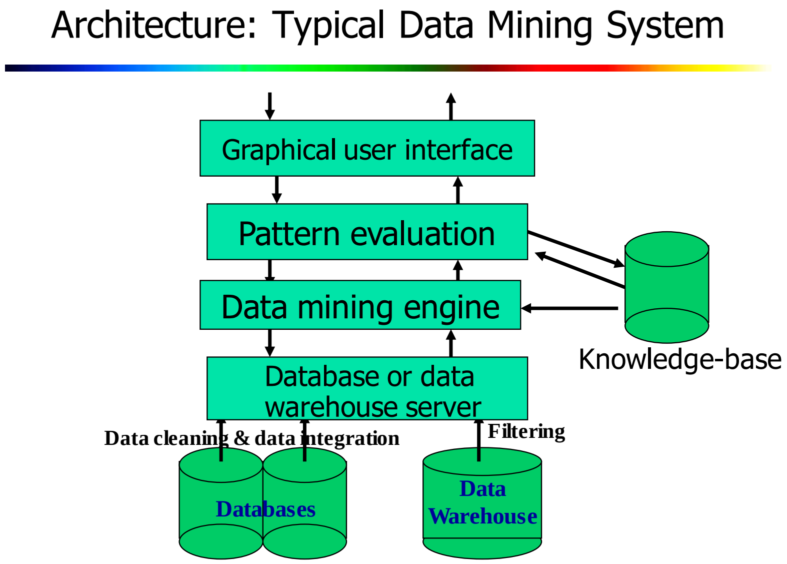


*Figure 2.1: Architecture of Data Mining*

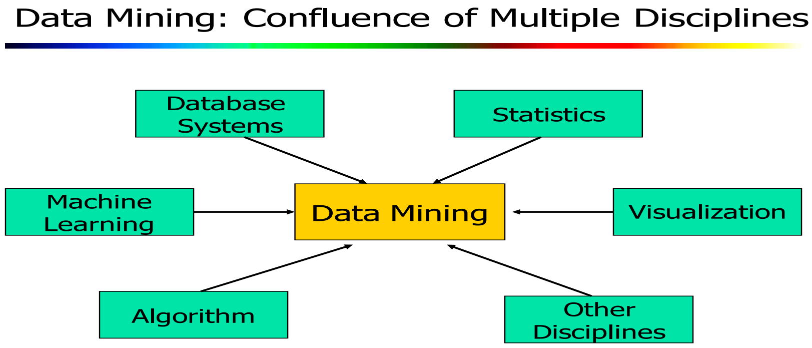


*Figure 2.2: Data Mining Confluence of Multiple Disciplines*

**Cryptographic methods for information sharing and privacy**

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties. The data may be distributed in two ways across different sites:

**Horizontal Partitioning:** In this case, the different sites may have different sets of records containing the same attributes.

**Vertical Partitioning:** In this case, the different sites may have different attributes of the same sets of records.

## Experimental results & decision

Partitioning in general, results in smaller, manageable data sizes, so indexes are built faster, queries run faster, more data can actually fit into memory, and so on.

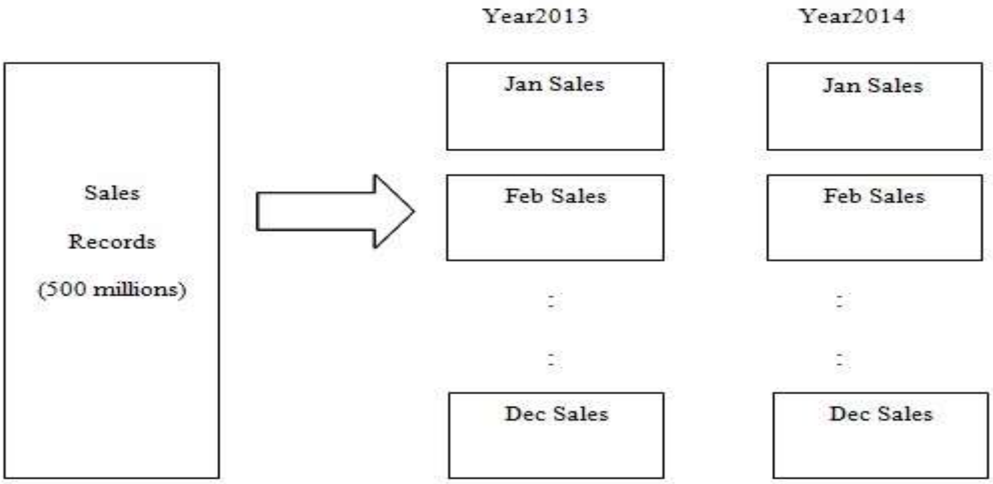
Partitioning a database improves performance and simplifies maintenance. By splitting a large table into smaller, individual tables, queries that access only a fraction of the data can run faster because there is less data to scan. Maintenance tasks, such as rebuilding indexes or backing up a table, can run more quickly. Partitioning can be achieved without splitting tables by physically putting tables on individual disk drives. Putting a table on one physical drive and related tables on a separate drive can improve query performance because, when queries that involve joins between the tables are run, multiple disk heads read data at the same time.

With Horizontal Partitioning, we keep the columns in a large table intact, but split the rows, again based on certain criteria so as to minimize querying across multiple partitions. a horizontally partitioned table might look like:

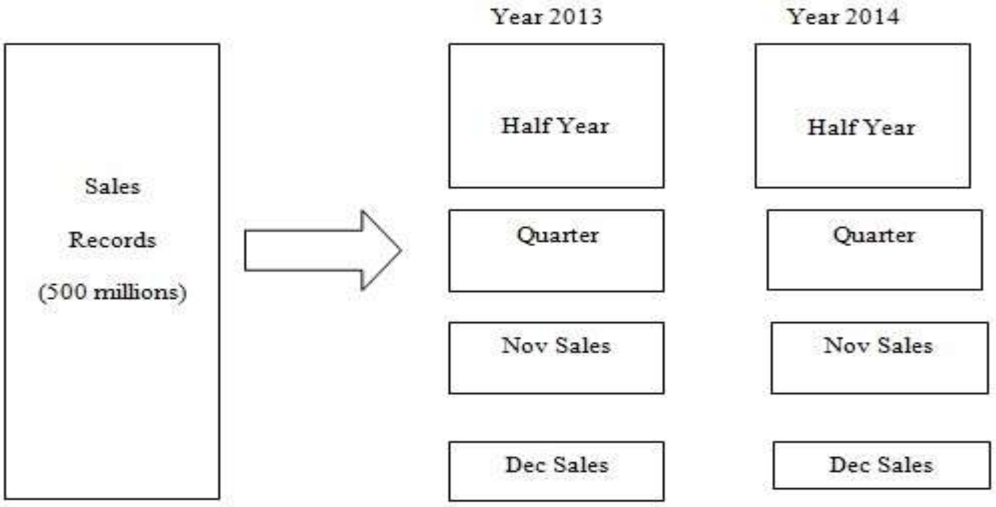
table_partition_1: n/k rows, 'm' columns
table_partition_2: n/k rows, 'm' columns
…..
.....
table_partition_k: n/k rows, 'm' columns

## Horizontally partitioned

1. Partition by time into equal segments is given



2. Partition by time into different sized segments



3. Partition on a different dimension, e.g. region
4. Partition by size of table

***Table 4.1(a): User_1***

| ID | Outlook | Temp | Humidity | Wind | Result |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |

***Table 4.1(b): User_2***

| ID | Outlook | Temp | Humidity | Wind | Result |
|---|---|---|---|---|---|
| 3 | overcast | hot | high | false | yes |
| 4 | rain | mild | high | false | yes |

**:**

**:**

***Table 4.1(f): User_7***

| ID | Outlook | Temp | Humidity | Wind | Result |
|---|---|---|---|---|---|
| 13 | overcast | hot | normal | false | yes |
| 14 | rain | mild | high | true | no |

***Table 4.1 Accessed Data by User***

**ID3 Algorithm**

```
ID3(Examples, Target_attribute, Attributes)
    1. Create a Root node for the tree
    2. If all Examples are positive, Return the single-node tree Root, with label = +
    3. If all Examples are negative, Return the single-node tree Root, with label = -
    4. If Attributes is empty, Return the single-node tree Root, with label = most common value of Target_attribute
       in Examples
    5. Otherwise Begin
        ▪ A ← the attribute from Attributes that best classifies Examples
        ▪ The decision attribute for Root← A
        ▪ For each possible value, v1, of A,
            ▪ Add a new tree branch below Root, corresponding to the test A= v1
            ▪ Let Examples v1 be the subset of Examples that have value v1 for A
            ▪ If Examples v1 is empty
                ▪ Then below this new branch add a leaf node with label = most common value of
                  Target_attribute in Examples
                ▪ Else below this new branch add the subtree
                  ID3(Examples,Target_attribute,Attributes       {A} )
    6. End
    7. Return Root
```

By using ID3 algorithm to mine on the union of datasets, we can obtain the public decision tree, while each party's private information are all revealed.

For preserve each party's private data two notions are used.

One is Privacy-Preserving Decision Tree, which is stored at the miner site. The semihonest miner only knows the basic structure of the tree, and which site is responsible for the decision made at each node (i.e., only know which site possesses the attribute to make decision at the node, while without the knowledge of which attribute it is and what attribute values it has);

The other is Constrain Set {AX1, BX1}, it means that this path which is form the root node to the present node (the node with the value of BX1) has determined by those attributes in the Constrain Set. When beginning to build tree, all parties will send the numbers of local attribute to miner, and the Constrain Set is initialized as {}, as Constrain Set of the present node becomes full, i.e. {AX1, BX1, AX2, BX2}, it means X is empty , the next node should be leaf node, which with the class attribute value c, assigned to most transactions with the certain transaction IDs.

When the miner creates a root node, it sends signal to all parties. Each party obtains the local best prediction attribute Xi by information gain measurement, then sends the attribute serial number Xi and information entropy to the miner by Protocol for Comparing Information Without Leaking (PCIWL), which ensures that no original information would be revealed at miner site or any other parties. The miner applies Protocol for Comparing Information Without Leaking (PCIWL) to get the maximum as the global best prediction attribution, while he doesn't know the which attribute it is and what attribute values it has, he just has the knowledge that which site possesses that attribute and its' serial number, e.g., as it is shown in above figure, the minor creates a root node AX1, that isUser_1 has the information at that node, and the first attribute possessed by User_1 is the best prediction attribution. At the same time, the minor set {AX1} as Constraint set of the present node. When creating the next node, whether it's a leaf node or internal node, the steps is used.

Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes

$$\text{Entropy}(S) = \sum -p(I) \log2\ p(I)$$

Gain(S, A) is information gain of example set S on attribute A is defined as

$$\text{Gain}(S, A) = \text{Entropy}(S) - □\ ((|S_v| / |S|) * \text{Entropy}(S_v))$$

We study and analysis the multi party privacy preserving decision trees for horizontally partitioned data that when building decision tree, the control is passing from site to site, except token party has the knowledge of best prediction attribute of the present node, other party even the miner doesn't know any relevant information. When classifying, the miner only knows the path of classifying process, i.e., which site handles the classifying in every step, while the information of which attribute is used to classify and values of transaction records in every party is protected. Entropy is calculated as

***Table 4.2: Entropy Estimation***

| **Entropy** | **value** |
|---|---|
| Entropy(S) | 0.940 |
| Entropy($S_{weak}$) | 0.811 |
| Entropy($S_{strong}$) | **1.00** |

Gain is calculated as

***Table 4.3: Gain Estimation***

| ***Gain(S, Outlook)*** | ***0.246*** |
|---|---|
| Gain(S, Temperature) | 0.029 |
| Gain(S, Humidity) | 0.151 |
| Gain(S, Wind) | 0.048 |
| Gain($S_{sunny}$, Humidity) | 0.970 |
| Gain($S_{sunny}$, Temperature) | 0.570 |
| Gain($S_{sunny}$, Wind) | 0.019 |

INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

## Conclusion

In this research work, we discussed methods for distributed privacy-preserving mining, and the methods for handling horizontally partitioned data. The primary contribution of this work is to propose a multi party privacy preserving decision tree for horizontally partitioned data by using ID3 algorithm. This has particular relevance to privacy-sensitive searches, particularly top-k queries, and meshes well with privacy policies. There remain many open problems in developing secure solutions based on efficient non secure query processing algorithms. Further, this work has shown that there is a trade-off between efficiency and the amount of information that is disclosed. It is worthwhile to explore whether one could have a suite of algorithms (or a configurable algorithm) so that applications can choose the goal they want to optimize. Finally this gives the best privacy preserving, efficiency and accuracy.

## References

1. Grljevic O,Bosnjak Z,Mekovrc R., “Privacy Preserving in Data Mining- Experimental research on SMEs data”, 2011 IEEE International Symposium on Privacy Preserving, Vol.4, pp. 477-481, October 2011
2. K. Murat and Chris Clifton., “Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data”, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp.1026-1037, September 2004
3. Newton E., Sweeney L., Malin B., “Preserving Privacy by De-identifying Facial Images”, IEEE Transactions on Knowledge and Data Engineering, IEEE TKDE, pp.1013-1021, February 2005.
4. Li, Y., Chen, M., Li, Q. Zhang, W., ” Enabling Multi-Level Trust in Privacy Preserving Data Mining ”, IEEE Transactions on Knowledge and Data Engineering ,Volume: PP Issue:99, page 1-11,09 ,June 2011
5. Zhe Jia, Lei Pang, Shoushan Luo, Yang Xin.: Miao Zhang, "Research on Distributed Privacy- Preserving Data Mining", JCIT: Journal of Convergence Information Technology, Vol. 7, No. 1, pp. 356-367, 2012
6. Y. Lindell and B. Pinkas., “Privacy preserving data mining”, In Advances in Cryptology, volume 1880 of Lecture Notes in Computer Science, pp. 36–53, Springer-Verlag, 2000
7. J vaidya, C Clifton., “Privacy Preserving Kth Element Score over Vertically Partitioned Data”, IEEE Transaction on Knowledge and Data Engineering, Vol.21,No.2, pp.253-258, February 2009
8. Sumana M and Dr Hareesh K S., “Anonymity: An Assessment and Perspective in Privacy Preserving Data Mining”, International Journal of Computer Applications Record 6(10), pp.1–5, September 2010.
9. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis., “State-of-the-art in privacy preserving data mining”, ACM SIGMOD Record, 3(1), pp. 50–57, March 2004.
10. Nan Zhang ,Wei Zhao, ”Privacy Protection against Malicious Adversaries in Distributed Information Sharing Systems”, IEEE Transaction on Knowledge and Data Engineering, Vol.2.0, No.8, pp. 1028-1033, August 2008
11. H. Yu, X. Jiang, and J. Vaidya.: “Privacy-preserving svm using nonlinear kernels on horizontally partitioned data”, In Proceedings of the ACM Symposium on Applied Computing, pp. 603–610, Dijon, France, 2006.
12. Bawa M., Bayardo R. J., Agrawal and J Vaidya, “Privacy-Preserving Indexing of Documents on the Network”, VLDB Conference, pp.23-30, 2003.
13. J. Vaidya and C. Clifton.: “Secure set intersection cardinality with application to association rule mining”, Journal of Computer Security, 13(4), pp. 24-31, July 2005.
14. Jaideep Vaidya ,Chris Clifton, “Privacy-Preserving decision trees over Vertically Partitioned Data, Proceedings of the 19th annual IFIP working conference on Data and Applications Security, pp.139-152, August 2005,
15. Clifton C.,Kantarcioglou M., LinX., ZhuM.: “Tools for privacy-preserving distributed data mining”, ACM SIGKDD Explorations, 4(2),pp. 342-348, 2002.
16. Du W., Atallah M.: “Secure Multi-party Computation, A Review and Open Problems”, CERIAS Tech. Report 2001-51, pp.1-51, Purdue University, 2001.
17. Lindell Y., Pinkas B.: “Privacy-Preserving Data Mining”, CRYPTO, pp.176-190, 2000.
18. Kantarcioglu M., Vaidya J.: “Privacy-Preserving Naive Bayes Classifier for Horizontally Partitioned Data” IEEE Workshop on Privacy-Preserving Data Mining, pp.256-261, 2003
19. Yu H., Jiang X., Vaidya J.: “Privacy-Preserving SVM using nonlinear Kernels on Horizontally Partitioned Data”, SAC Conference, pp.312-317, 2006
20. Murat Kantarcioglu , Chris Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data”, IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.9, pp.1026-1037, September 2004