Available online at www.ijiere.com

International Journal of Innovative and Emerging Research In Engineering

e-ISSN: 2394 - 3343 p-ISSN: 2394 - 5494

Privacy Preserving Data Mining With Classification And Encryption Methods

Sarvaiya Sukhdev and Hemant Vasava

Final year ME scholar, Comp. Dept. Birla Vishvakarma Mahavidyala and affiliated to Gujarat Technological University Ahmedabad, India

Assistant professor at Birla vishvakarma Mahavidyalaya in Comp. Dept. and affiliated to Gujarat Technological University Ahmedabad, India

ABSTRACT:

Data mining can be performed with four different directions Association, Clustering, Classification and the Bayesian formula. In the recent years, many privacy preserving techniques are developed. These techniques are used to provide the privacy to the data. Here privacy related various information is gathered with their brief introduction. Here mainly focus is on the classification decision tree and the various encryption methods. In the cryptography both symmetric and asymmetric encryption algorithms are described here. Among them any of the encryption algorithms can be used in the privacy preserving data mining. So here both Decision tree algorithms and some encryption methods are surveyed.

Keywords: Data mining, classification, Cryptography

I. INTRODUCTION

Data mining is the process of discovering interesting pattern or knowledge from the large amount of data. It used in the field of business intelligence, Web search, scientific discovery and digital library etc. [1]. Data mining can be performed with four different directions Association, Clustering, Classification and the Naïve Bayesian Data mining. In the recent years, many privacy preserving techniques are developed. These techniques are used to provide the privacy to the data. Here privacy related various information is gathered with their brief introduction. Some of the data mining issues are mentioned below [4].

a) There is not any unifying theory in data mining.
b) Increase the dimensionality and high speed streaming of data.
c) Mining of sequence and time series data.
d) Obtain complex knowledge from complex database.
e) Network configuration
f) Data mining in distributed environment or multi agent data.
g) Biological and environmental problem related data mining.
h) Issue of process related problems.
i) Privacy, security and integration of data.
j) Data mining dealing with non-static, unbalanced and cost sensitive data.

A. DATA MINING

We have seen that a data mining is the process of the knowledge discovery from data (KDD).To discover that knowledge following steps are included in that procedure.(1) Data preprocessing :It performs the basic operations like data cleaning, Data integration and selection process. (2) Data Transformation: In this step data are transformed to the appropriate form for the further mining task. (3)Data Mining: This is an essential process of Mining where Association, Clustering or Classification like intelligent methods are used for the mining. (4)Evaluation and presentation: Presenting the knowledge in an easy to understand fashion. [1]

Here various data mining techniques are mentioned below.

a. Clustering
Clustering is a one of the data mining technique in which data similarities are placed in one group. This is a process similar to the data segmentation. It used in the field of life science, medical science and engineering and so on. Main advantages of the clustering is interesting patterns and knowledge can be found directly from the large amount of data without any prior knowledge of the cluster. There are various clustering algorithms like hierarchical clustering, K mean and Density based clustering algorithms.

b. Classification
Classification is also one of the data mining technique which widely used in various data mining domain. This is a supervised learning technique in data mining. Various phases in the classification like learning phase, testing phase and the application phase are there. Various models are existed for the classification like decision tree, Naïve Bayesian classification and the support vector machine etc.

c. Association rule:
Association also plays an important role in data mining which used for discovering interesting relation between variables in large database. This data mining technique works based on the frequent patterns and the basis of their support values. Various algorithm are existed for the association rule mining like A-priory algorithm, Context based association rule mining, node set based algorithm, sequential pattern mining, OPUS search etc.

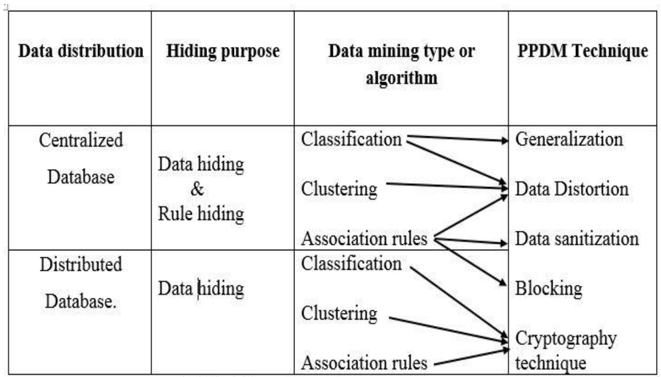
B. PRIVACY PRESERVING TECHNIQUES:

The main goal of the privacy preserving is to provide the ultimate security to our sensitive information. Various privacy preserving techniques are enlisted below [2].

- Randomization
- Anonymization
- Secure multi party computation
- Sequential pattern hiding.

C. PRIACY PRESERVING DATA MINING TECHNIQUES

Other Data modification or Data Perturbation techniques are existed like noise addition, Data swapping, aggregation, suppression, sanitation, data hiding and data preprocessing etc. [2]. For the two type of the data distribution, centralized and distributed database with their data mining type or algorithm [4]. Various PPDM techniques are shown here



Data distribution	Hiding purpose	Data mining type or algorithm	PPDM Technique
Centralized Database	Data hiding & Rule hiding	Classification Clustering Association rules	Generalization Data Distortion Data sanitization
Distributed Database.	Data hiding	Classification Clustering Association rules	Blocking Cryptography technique

PPDM Algorithms taxonomy

Figure 1. PPDM Algorithm Taxonomy

Here some Privacy preserving are listed with their advantages and disadvantages [7].

a. PERTURBATION TECHNIQUE:
This technique used for the privacy preservation in which data are perturbed
But this cannot reconstruct the original data and also not good for the large data.

b. CONDENSATION TECHNIQUE:
Instead of perturbed data, it works on the pseudo data. So it provide better privacy preservation than the techniques which use simply data modification on original data.
But it does not give longer effect on data mining. Because it has the same format as the original data.

c. CRYPTOGRAPHIC TECHNIQUE:
It performs encryption of the sensitive data. There is also proper toolset for algorithm in the field of the data mining.

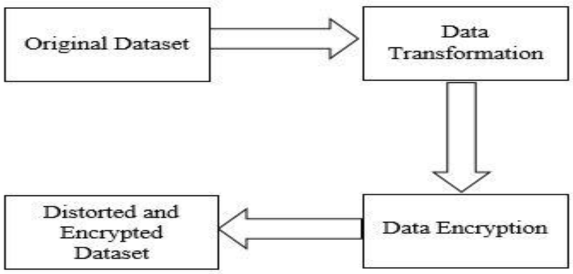
But this technique is difficult to scale when more parties are involved and also not good for large database.

d. BLOCKING BASED TECHNIQUE:

In this technique to provide privacy to the individual it replaces the unknown values to the sensitive transaction. Reconstruction of the original data are quite difficult.

D. COMBINE STRATEGY FOR THE PRIVACY PRESERVING DATA MINING.

In the combine strategy multiple strategies are used to obtain any privacy preserving methodology. In the given figure combine strategy is shown based on the data transformation and data encryption techniques. Due to the combining this various techniques robust security can be obtained. Sometime privacy preserving techniques may have some disadvantages or some limitations but that can be overcome in combine strategy. So security result would be more effective of combine strategy than a single privacy preserving techniques used [3]. In the given figure original data are transformed after that transformed data are encrypted. So here data transformation and data encryption methods are used in this combine strategy.



PPDM Methodology

E. CLASSIFICATION:

Classification is a kind of process which performs a data analysis that extracts important data classes. Entire classification technique is performed in two steps. First learning step in that step any classification algorithms is apply on the dataset and build the classifier. In second step those classifier are used to classify that dataset. Various typical classification models are as below [1].

a. DECISION TREE:

This is one of the classification model used to classify data in the tree structure. In that structure each internal node represent a test on that attribute, each branch represent the outcome of the test and leaf node represent the class of the data. A decision tree can be easily converted to a classification rule [1] [7].

b. NAÏVE BAYESIAN CLASSIFICATION

In this classification model, Bayesian rule based on the posterior probability is used. In this Bayesian theorem one assumption is taken as that the effect of the values of the class attribute is independent than the values of the other attributes [1].

c. SUPPORT VECTOR MACHINE

Support vector machine (SVM) also plays an important role in the classification. For transforming the original data to higher dimension, SVM uses linear mapping. Using this new dimension SVM search for the linear optimal separating hyper-plane which is the decision boundary indicates the separation of tuples of the class from another [1].

F. CLASSIFICATION AGLGORITHMS:

Classification algorithms also plays an important role in the data mining for analyzing the data. There are various classification algorithms.

a. ID3
b. C4.5
c. CART
d. K-NN
e. Naïve Bayes etc.

Some of these ID3, C4.5 and the CART algorithms are explained here.

a. ID3 (Iterative dichotomier tree) Algorithms.

This is a kind of decision making tree algorithms used in the classification Data mining. It is based on Entropy and the Information gain. Basic idea behind use of the ID3 is to ask the question whose answer provide most information. For the given dataset ID3 chose that attribute who have the maximum information gain [5] [6].
The information Gain is calculated as
Entropy(S) =-$\sum_{i=1}^{n}(P_i \log_2 P_i)$
Where, P_i is the probability that an arbitrary tuple in S belongs to class C_i
Gain (A) = Entropy (S) - Info (S)
S_i = {S_1, S2……..S_n} partition of S according to value of attribute A.

b. C4**.5 :**
This algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible. This algorithms includes following steps.
1. Choose the attribute as the root node.
2. Create branch for each value of that attribute
3. Split cases according to branches.
4. Repeat process for each branch until all cases in the branch have the same class [5] [6].

c. CART
This algorithms measure the impurity of the dataset D .The data Partition or the training tuples as
Gini (Dataset D) =1- $\sum_{i=1}^{n} \boldsymbol{P_i^2}$ (Where P for probability of ith Class attribute) In the CART algorithm, It performs binary split of the attribute values.
When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on a partitions D into D1and D2.The Gini index of D given that partitioning is
GiniA (D) = |D1|/|D| Gini (D1) + |D2|/|D| Gini (D2)
The reduction in impurity that would be incurred by a binary split on a discrete-valued or continuous-valued attribute A is
ΔGini (A) =Gini (D)-GiniA (D)
There are many advantages of the CART algorithms as followings.

- It is Non parametric (no probabilistic assumption)
- It can use any combination of the continuous /discrete variables
- It handles missing values automatically (using surrogate splits)
- In the compare of the other decision making tree like ID3 and C4.5, It has more accuracy and less time complexity [5] [6].
- From the described all three algorithms above in respect of accuracy and time complexity CART is performs better than other two algorithms [6].

G. CRYPTOGRAPHY

The concept of the cryptography includes

Plaintext: Refers to the source message or the original message before it encrypted.

Cypher text: Receiver of destination message refers to the Target message which is processed with encryption algorithm with suitable encryption and decryption keys.

Encryption Algorithm: The algorithm which gives the encrypted plain text called cipher text.

Decryption Algorithm: The algorithm which gives the decrypted cipher text and which is plain text.

Encryption or Decryption Key: The key which used in encryption or decryption algorithm. It may be of various size, numeric or string etc [8].

The modern field of the Cryptography can be divided into several areas of study. The Chief ones are discussed here.

a. Symmetric or Secret key cryptography
b. Asymmetric or public key Cryptography.

a. Symmetric Or Secret Key Cryptography
It refers to an encryption method in which both sender and receiver shares the same Key. These Ciphers are implemented as either block cipher or stream cipher. Here encryption key and decryption key both are same. These techniques are simpler and faster but their main drawback is that two parties must somehow exchange the key in a secure way. There are various symmetric algorithms available algorithms existed for

the encryption like the Caesar cipher, Play fair, rail fence and other stream cipher and block ciphers encryption methods are includes. Her in our proposed algorithm we have use Simple Caesar cipher and AES encryption algorithm for the privacy preservation of the dataset [8].

b. Asymmetric or Public Key Cryptography

Public key can be decrypted only with the corresponding private key. Public key encryption (also called asymmetric encryption) involves the pair of keys, a public and a private key. Which is associated with the entity. Each public key is published and corresponding private key is kept secret like the RSA and diffie Hellman algorithms

II. CONCLUSION

This survey paper includes the survey of the Data mining, various data mining algorithms, privacy preserving techniques, PPDM taxonomy, Classification algorithm and various decision tree algorithms are given. Along these all techniques, may be some of the techniques have some limitation or disadvantages. To overcome these limitation or disadvantages combined strategy is used. From our survey analysis we can conclude that which one is the best algorithms in the classification and decision tree. For robust privacy preservation Combine strategy may be used. For the robust privacy of the data using encryption method, some iterative or multiple key encryption algorithm should be used.

REFERENCES

[1] LEI XU, CHUNXIAO JIANG, (Member, IEEE), JIAN WANG, (Member, IEEE), JIAN YUAN, (Member, IEEE), AND YONG REN, (Member, IEEE*)," Information Security in Big Data: Privacy and Data Mining*", Page No 14-17, VOLUME 2, 2169-3536 , 2014

[2] Tamianna Kachwala and sweita parmar,"*An approach for preserving Privacy in Data Mining*" , Volume 4, Issue 9, Page No 1-4, September 2014 ISSN: 2277 128X.

[3] Santosh Kumar bhandare ,"*Data Transformation and Encryption Based Privacy Preserving Data Mining System*", , IJARCSSE Volume 4, Issue 7, Page No 3-4, July 2014 ISSN: 2

[4] K. Sriniivasa Rao & B. Srinivasa Rao," *An Insight in to Privacy Preserving Data Mining Methods*", CSEA Vol. 1, No. 3, Page No 1-3 , July-August 2013

[5] Monika Gupta," A Survey on Privacy Preserving Data Mining Using Decision Tree Algorithms", IJRETM ISSN 2347-7539 Volume: 02 Issue: 04 ,Page No 1-4 ,July-2014

[6] Subrata pramanik,Md.Rashedul Islam ,Md.Jamal Uddin, "*Pattern Extraction ,classification and comparision between attribute selection measures* ", International Journal of Computer Science and Information Technologies, Vol. 1 (5) ISSN:0975-9646 , Page No 2-5 ,2010

[7] Sweta Taneja,shashank Khannia,sugandha tilwalia,ankita," *A Hybrid C- Tree Algorithm for Privacy Preserving Data Mining*", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-4, Issue-ICCIN-2K14, Page No 2-3 ,March 2014

[8] Saranya k ,Mohanapriya r and Udhayan j "A Review on Symmetric Key Encryption Techniques in Cryptography" International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 3, Page No-1-6, March 2014