Secure Sum based Privacy Preservation Association Rule Mining on Horizontally Partitioned Data

Bhawani Singh Rathore
Student,Mtech,6th Semester
Branch C.S.E
UIT
Barkatullah University

Anju Singh
Assistant Professor
Department of CS & IT
UTD
Barkatullah University

Divakar Singh
H.O.D.
Department of C.S.E
UIT
Barkatullah University

ABSTRACT
The method of perturbation has been basically studied for the privacy preserving data mining. In this technique, from a known distribution random noise is combined to the private data before forwarding the data to the data miner. Consequently, the data miner constructs again a presumption to the original distribution of data from the perturbed data and the reconstructed distribution is used for the purposes of data mining. The goal of privacy preserving data mining researchers is to introduce techniques of data mining which could be implemented on the databases without break the privacy of the persons. Techniques of Privacy preserving for several models of data mining have been suggested, originally for the classification on the organized data then for association rules in the distributed area. This paper suggested a solution for the computing the data mining classification algorithm for the horizontally partitioned data privately without revealing any information related to the sources or data. The given method (PPDM) integrates the benefits of the RSA public key cryptographic system and homomorphic scheme of encryption.

Keywords
Horizontally Partitioned Dataset, Secure Sum, Privacy Preservation, Association Rules.

1. INTRODUCTION
Currently these types of databases are spread around all over the world. From various locations distributed data should be gathered in to the data warehouse, so that there is a need for a safe transmission of data and managing privacy. The data to be transmitted may include information that may be private to the individual or information of organization that should be safe as present in paper [1].

Privacy Preserving Data Mining (PPDM) is becoming a popular research area to address various privacy issues. Perturbation techniques and their techniques of privacy protection have been disapproved due to few methods may acquires personal information from the step of reconstruction [9]. Separate to the original noise additive method in [3], many distinctive perturbation methods have been proposed. There are several efficient, partially and number of totally homomorphic, but less effective cryptosystems. Although a cryptosystem which is by accident homomorphic can be matter to attacks on the basis of this, if cured carefully homomorphism could also be used to perform computations securely.

In this paper, RSA public key cryptosystem and homomorphic encryption are used to develop a reliable privacy-preserving data mining technique for horizontally partitioned data. Homomorphic encryption is a type of encryption method which lets specific kind of computations to be carried out on cipher text and get an encrypted result that decrypted combines the outcome of operations done on plaintext. The enormous growth of the Internet and its undemanding access by common man created opportunities for combined computations by numerous parties.

All the candidate parties for the benefit of their joint profits needs to calculate the normal function of their inputs but simultaneously they are concerned about their data's privacy. This matter of the security of information is known as Secure Multi- Party Computation. This matter has two main objectives; first one is privacy of the person's data inputs and second is accuracy of result. Basically two types of models are explained here for analyzing the Secure Multi-Party Computation issues. Optimal model of Secure Multi-Party [2] Computation uses a Trusted Third Party apart from the participating parties.

As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment. Data may be partitioned in various manners like vertical, horizontal and mixed. In the horizontal partitioning of the data, every fragment includes of a subset of records of R as relation and in vertical partitioning of the data, every fragment includes of a subset of the attributes of relation R.

The method of partitioning is mixed fragmentation in which data is partitioned first horizontally and then every partitioned fragment is again partitioned into vertical fragments or vice versa [4].

2. PRIVACY PRESERVATION IN ASSOCIATION RULE MINING
Many researchers proposed many methods for privacypreserving association rule mining for both centralizedand distributed databases. The state of the art in the environments of techniques for privacy preserving data mining is discussed by the authors in [4]. This paper also describes the different dimensions of preserving data mining techniques such as data distribution, data modification technique, data mining algorithms, data or rule hiding and approaches for privacy preserving datamining techniques.

Association rule mining is continuously getting more and more attention among data mining techniques to the researchers to explore correlations between items or item sets. These rules can be analyzed to make strategic decisions to improve the

performance of the business or quality of the organization service and so on.Association rule mining was introduced in paper [4].

An association rule may be explained as below.Let I = {i1,i2, …, im} can be a set of the attributes known as items. This item set X includes of one or many items.

Let DB = {t1, t2, … , tn} can be a database including n number of the Boolean transactions and every transaction ti includes the items supported by the ith transaction. One item set X is marked to be as frequent if the number of transactions that are supporting this item set is greater than or equal to user described the minimum support threshold or else it is called to be an infrequent. An association rule

is a conclusion of the form like X→ Y in which X and Y are separate subsets of I, and X is known as the antecedent and Y is known as a consequent. An association rule X→Y is termed as to be a strong association rule only if its confidence is more than or equal to the user describing minimum level of confidence.

Many researchers proposed various methods for privacypreserving association rule mining. It is for both centralized anddistributed databases. The several methodologies likeperturbation, randomization, cryptography and heuristicmethods are suggested in this paper to detect the privacypreserving association rule mining in the vertically and horizontally partitioned databases.

Among many different techniques cryptography technique is one of the very popular and heavily used technique to apply for horizontal, vertical and mixed mode partitioned dataset. It provides accurate and effective solution. It provides informational accuracy to users and at the same time privacy constraints are satisfied.

3. DATA PARTITION

The way that data partitioned is one of most importantfactor in distributed data mining. Majority of algorithms aredesigned and developed on the concept of data partitioning. Generally, there are two types of data partitioning, vertical partitioning and horizontal partitioning. In vertical partitioning the available data are stored at different geographic locations, for example suppose that in a data mining process data need to collect different data such as financial, medical, insurance, hospital, school and housing data about different person who resides in the city.

3.1 Horizontally Partitioned

Horizontal partitioning divides the whole database into the number of small database according to the row splitting. So that the execution of query will be very fast as well as it will be able to provide more privacy to the portioned database. Horizontally partitioned data can be used where every fragment includes a subset of records of R as relation. According to paper [5] [6] [7], horizontal partitioning method break a table into various tables.

In this tables have been partitioned in a pattern like the query references are done by the use of less number of tables or enormous UNION queries are used to combine the tables apparently at the time of query which can influence the performance. For example suppose that in a data mining project it is needed to investigate the effects of a drug on those patients who are having special illness. Specially in order to increase various samples it is needed to obtain the same information about this issue from different medical centers. In such settings it is said that the data are partitioned horizontally.

3.2 Vertically Partitioned

Vertical portioning is a technique which divides the complete dataset into a number of small databases according to the column. So that portioned database does not contain any of duplicate data. There are mainly two types of vertical database normalized and row splitting. The data may be break into the set of small files that are physical, each file is consists of the subset of original relation, the relation is database transaction which basically needs the subsets of given attributes.

In vertical partitioning the data available about a set of same entities are placed in different locations, for example suppose that in a data mining process it is needed to collect different data such as financial, medical, insurance and housing data about different people resident in a city. In this process it should gather different data about a set of same entities, i.e. those people in that city, from the servers of different institutions such as medical institutions, government servers, municipalities, banks and so on [8].

4. CRYPTOGRAPHIC METHODS

The cryptographic method to PPDM supposes that the data is recorded at many private parties, those are ready to explore the outcome of some data mining calculations done in combination over their data. The parties reserved in the protocol of cryptographic, that is they communicate messages encrypted to create few operations effective where as others computationally difficult. As a result of this, they “blindly” execute their algorithm of data mining.

Cryptography is a technique through which sensitive or critical information could get encrypted. It is a very effective way to preserve the data. In paper [9], authors introduced cryptographic technique that is very useful and effective because it provides security and safety of sensitive attributes. There are different algorithms of cryptography available .But this method has many disadvantages. It fails to protect the output of computation. It helps in preventing various privacy leakages in computation. The algorithm does not give fruitful results when it talks about more parties. It is hard to use the algorithm for huge databases.

Final data mining result may break the privacy of individual’s record. Oblivious transfer is a basic protocol that is the main building block of secure computation. It seems very strange at the first look, but the role it plays in secure computation should become clear later. Computationally intensive operations in oblivious transfer is often in secure protocols, and is repeated many times. Each invocation of oblivious transfer mainly requires a fix number of invocations of trapdoor permutations (i.e. public-key operations,). It is always feasible and effective to decrease the amortized overhead of oblivious transfer to one exponentiations per a logarithmic number of oblivious transfers, even for the case of malicious adversaries [10].

4.1 Secure Sum

Secure sum calculation issue of Secure Multi-Party Computation may be defined as: How multiple parties can compute the sum of their input values without disclosing definite values to one another. Secure sum can occupation as to implement for the Secure MultiParty Computation solutions in the privacy preserving dispersed data mining problems [11]. It is suggested a novel Rk-secure sum protocols with more security in case a group of the parties join together and want to know the private data of some other party

Secure sum [11] is applicable only for two parties for providing the security. In this protocol one party send the

partial support to the next party with adding their own random number then the last party will disclose the result.The method of secure sum is a method for combining results on distributed servers, in a way that the financial result which is the summation of local results will be acquired without appearing of any local results.

The secure sum has been used as one of the important methods in combining private results of sub algorithms. For example, [9] which have been used secure sum as the major result combining module. Although this famous algorithm has been used greatly in these fields, it has weaknesses.

For example it can refer to the collusion of two servers for accessing to the information of server that is between them. In this paper, by introducing improvement on the secure sum algorithm which is resistant against the collusion between two members will use it as the major combining module. The purpose in secure sum is calculating the summation of all distributed results without disclosure any of them.

4.2 Secure Multiparty Computation

All these methods are almost based on a unique encryption protocol called as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining made up of a set of protected sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure size of intersection, secure set union and scalar product [19].

Advantages

- Safe
- Secure
- Trust-worthy

Disadvantages

- Communication complexity grows exponentially with n

5. LITERATURE REVIEW

Privacy-Preserving Data Mining (PPDM) [12] is a newresearch area that investigates how the privacy of datacan be maintained either before or after applying DataMining (DM) techniques on the data. Previous task in the privacy preserving data mining is depends on the two methods. First one is the target to prevent the privacy of the customer by perturbing the data values. The basic assumption of this method is the perturbed data never disclose the private information, and hence it is ‘‘safe’’ to be use for the data mining.

Algorithms of data mining which partition the data into various subsets have been introduced. In specific, operation in parallel data mining can be appropriate. Parallel data mining algorithms can also work as initial point. Algorithms have been suggested for the distributed data mining. A method is suggested for the horizontally partitioned data in paper, and currently work has presented privacy in this scheme. Distributed classification has also been presented.

An approach of meta-learning has been introduced which uses classifiers prepared at various sites to evolve a global classifier. This may prevent the separate entities, but it still remains to be presented that the each classifiers do not reveal personal information. Current work has been addressed classification by the use of Bayesian Networks in the vertically partitioned data, and the circumstances in which the distribution is interesting according to what is known.

Although, those algorithms do not pay full attention to data mining results, which may lead to sensitive rules leakages. While some algorithms are designed for preserving the rule such as with sensitive information, it may degrade the accuracy of other non-sensitive rules.

Thus, further investigation, focusing on combining data and rule hiding, may be beneficial, specifically, when taking into account the interactive impact of sensitive and non-sensitive rules. Fourth, although many machine learning methods have been used for classification, clustering, and other data mining tasks (e.g., diagnose, prediction, optimization), currently only the association rules method has been predominately used for classification.

It would be interesting to see how to extend the current technique and practice into other problem domains or data mining tasks. Furthermore, it is important to find the privacy preserving technique that is independent of data mining task so that after applying privacy preserving technique a database can be released without being constrained to the original task.

Here in this research in protected distributed computation, which takes place like a large body of research in the concept of cryptography, obtained great results. It provided non trusting parties can connects to compute functions having distinct inputs at the time ensuring that no party seeks anything but provides output as a function.

These results displayed through generic constructions which are applicable to any function and offer an efficient presentation like a circuit. Authors explain their results, focus on their efficiency, and analysis their relevance to privacy preserving calculation of data mining algorithms. Here they displayed some examples of protective computation of the data mining algorithms which implements these generic constructions.

6. PROPOSED WORK

Data Mining plays very important aspect in various applications. Proposed method of privacy preservation has become more important just because of it’s data utility. The architecture is describing fig 1.

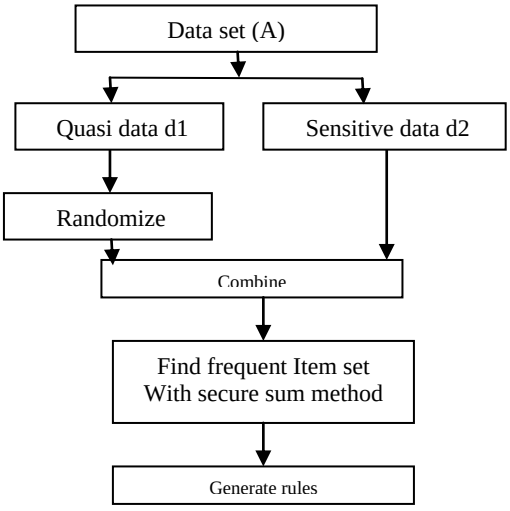


Fig 1: Architecture of proposed work

Figure 2 show the algorithm of the proposed work.

1. D←read dataset
2. D1, D2, D3and D4 are dataset of P1, P2, P3 and P4 party respectively.
3. Each party identify qusi and non -qusi attribute .
4. For each party randomize(Di(:,1:noQ)) where 1<=i<=4 and noQ is number of qusi attribute
5. For i=1 to 4 repeat 6
6. For each single itemset belongs to Di calculate actual support(ACS i,j)// ith party jth itemset
7. For each itemset repeat 8 to10
8. Ps(j)=R+ ACS (1,j)*size(D1)
9. For i=2to4 repeat10
10. Ps(j)=Ps(j)+ ACS (i,j)*size(Di)
11. Total support ts(j)=Ps(j)-R/(SUM(size(Di))i=1,2,3,4)
12. Select those item whose ts>=minsup store them in T and corresponding suppprt in tempSupport
13. Frequentitem(1)=T, Support(1)= tempSupport
14. For k=2 to Noofattribute repeat 15 to 22
15. Temp=T;
16. T=[];
17. Combine (temp,k) // combining element of last frequent item set taking j at atime
18. j = 1:size(Combinations,1) Repeat 19 to 21
19. For each itemset belongs to Combine calculate actual support(ACS(j)
20. If ACS(j)>minsup
21. Add combine(j) to T, corresponding suppprt in tempSupport
22. Frequentitem(K)=T, Support(k)= tempSupport
23. For each frequentitemset calculate ActualConfidence(X→Y)=Support(XUY)/support(X)
24. If ActualConfidence>minConf then Rules(1)←X And Rules(2)←(Y)

Fig 2: Proposed Work

7. RESULT ANALYSIS

This section mainly concentrates on two parts. First part discusses about the data source and system on which results are calculated. In second part will concentrate on result section.

Part 1:

Considered dataset is taken from UCI. Heart Disease dataset [last] is taken for the experiment results. It is very know university which provides various dataset for research purpose.

Number of cities: four.

Name of Cities:

1. Cleveland
2. Hungary
3. Switzerland
4. VA Long Beach

Total Number of patients city-wise: 920.

1. 303
2. 294
3. 123
4. 200

There are total 76 Attributes out of which 10 sensitive attributes are taken.

System on which experiments are performed and evaluated is as follows:

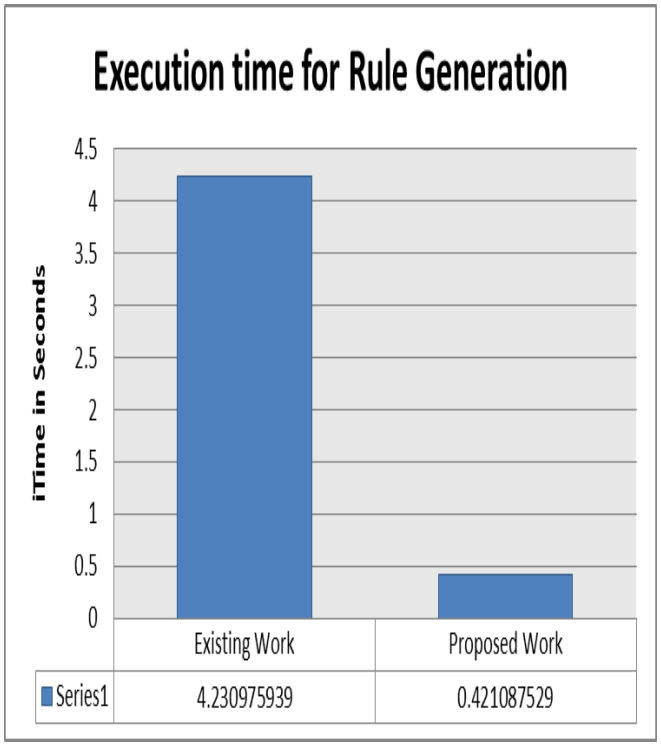
1. RAM: 4 GB
2. Processor: i3
3. Operating System: Windows 7 with 32 bits

Part 2:
There are two parameters on which result is evaluated.

a. Execution time

b. Privacy Preservation

Execution time : It is a quantity of time , measure in seconds. This time show the execution of algorithm and generate it's results. Execution result is shown in graph 1 and table I.



Graph 1: Execution Comparison of Proposed with Existing work.

Table 1: Execution time

Existing Work	Proposed Work
4.230976	0.421088

2. Dissimilarity Matrix

It is a matrix which shows the dissimilarity between original dataset in reference to the perturbed dataset. If it is bigger then

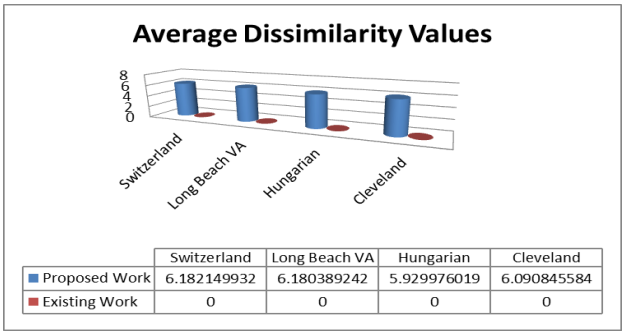
it show that the privacy is higher. If this value is lesser than it shows it provide lesser privacy level.

Table 2: show that propose work performs better on this parameter.

S. No.	Switzerland	Long Beach VA	Hungarian	Cleveland
1	6.069387968	6.033275859	6.013112914	5.69982275
2	6.286675008	6.199179499	5.885657432	6.438848921
3	5.919090814	6.179241731	5.903056278	6.440829945
4	5.985767907	6.096260182	5.964028777	5.964028777
5	6.324783651	6.383712864	5.910352569	6.23782713
6	6.180429569	6.337634025	5.863130772	6.071733917
7	6.479199249	6.110173019	5.952063881	6.00093838
8	6.353247837	6.191192501	5.929103526	6.00542175
9	6.19862371	6.24282062	5.894892597	6.068188927
10	6.024293609	6.030402125	5.984361447	5.980815348

where as there , all elements of dissimilarity matrix are 'ZERO' of existing work.

The comparison between the average dissimilarity values of proposed work with respect to existing work is shown in graph 2 along with it's individual values in table III.



Graph 2: comparison between the average dissimilarity values of proposed work with respect to existing work.

Table 3: Average dissimilarity values of proposed work with respect to existing work

	Switzerland	Long Beach VA	Hungarian	Cleveland
Proposed Work	6.182149932	6.180389242	5.929976019	6.090845584
Existing Work	0	0	0	0

8. CONCLUSION

The privacy preserving in many researches has been discussed as a main problem and some solutions have been suggested for it. In distributed data mining the problem of privacy preserving has become as a big problem too, which some solutions have been represented for it. Of course, each of represented solutions has weaknesses. This is an enhancement over the existing protocols that confirm the security for two works. Last session shows the proposed work is much better than the existing work on execution time and privacy maintenance.

9. REFERENCES

[1] Kiran P, S Sathish Kumar and Dr Kavya "A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining", An International Journal (ACIJ), Vol.3, No.2, March 2012.

[2] "Modified Distributed Rk Secure Sum Protocol", Jyotirmayee Rautaray, Raghvendra Kumar, Garima Bajpai, International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 3, March 2013.

[3] M tamer Ozsu Patrick Valduriez, Principles of Distributed Database Systems ,3 rd Edition.

[4] "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", N V Muthu lakshmi,Dr. K Sandhya Rani, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 3176 – 3182.

[5] Sugumar, Jayakumar, R., Rengarajan, C "Design a Secure Multi Site Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105. 2012.

[6] N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.

[7] N V Muthu lakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.

[8] "Distributed algorithm for privacy preserving data mining based on ID3 and improved secure sum", Ehsan Molaei, Hossein Vadiatizadeh, Amirmahdi mohammadighavam, Neda Rajabpour, Fatemeh ziasistani, International Journal of Advanced studies in Computer Science and Engineering IJASCSE, Volume 3, Issue 1, 2014.

[9] "A Review on Privacy Preserving Data Mining: Techniques and Research Challenges", Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2310-2315.

[10] "Implementation Of Cryptography For Privacy Preserving Data Mining", Anand Sharma and Vibha Ojha , International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010.

[11] R. Sheikh, B. Kumar and D. K. Mishra, "Changing Neighbors k- Secure Sum Protocol for Secure Multi-party Computation," Accepted for publication in the International Journal of Computer Science and Information Security, USA, Vol.7 No.1, pp. 239-243, Jan. 2010.

[12] Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le, "A Survey on Privacy Preserving Data Mining" ,in IEEE, 2009 First International Workshop on Database Technology and Applications.