ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Privacy Preservation in Data Mining by Hybridization of Partitioning on Medical Data

Madhan Subramaniam*
Asst. Professor,
Department of CSE,
Anna University of Technology – Tiruchirappalli,
Thirukkuvalai Campus, Tamilnadu, India
madhan444@yahoo.com

Senthil R
Programmer Analyst,
Infosys,
Chennai,
Tamilnadu, India

Bathmanaban L
Asst. Professor,
Department of ECE,
Anna University of Technology – Tiruchirappalli,
Thirukkuvalai Campus, Tamilnadu, India

Suhasini Manoharan
Asst.Professor, Department of CSE,
MIET Engineering College – Tiruchirappalli,
Tamilnadu, India

Abstract- Data mining is increasingly being used to improve the decision taking skills of the physician. As medical data is highly sensitive to personal information of human being, so it is desired to keep private. There are many approaches for classification which have been adapted for privacy preserving in medical data which are based on data separation technique for privacy preserving. There are two scenarios, one is centralized and other is distributed data where several approaches have been developed. In this paper, we propose architecture for privacy preserving in data mining by combining horizontal data distribution and vertical data distribution.

Keywords: Privacy, Security, Secure Multiparty Computation.

I. INTRODUCTION

Data mining is a technique which is useful to extract knowledge in knowledge discovery process from large databases. Medical domain has one of the most leading data set, where we use this technique for extracting or mining useful knowledge. Privacy preserving in medical domain is one of the challenging and emerging task and so it is attracted by many researchers and practitioners to do work in the area [8]. To preserve the privacy on medical data, classification technique is one of the attractive techniques in data mining because it generates rules. Medical data is so sensitive and it has patient's personal information so, privacy is the main issue and it can be solved by using classification technique. This paper gives the classification technique which is useful in both scenarios, centralized as well as distributed because in some cases we have to work in centralized environment as on the other hand we have to work in distributed environment. There are many approaches which have been adopted in privacy preserving data mining. We can classify them based on the following dimensions: data distribution, data modification, data mining algorithm, data or rule hiding and privacy preserving [8].

The objective of this paper is to apply data distribution technique to preserve privacy inclassification of medical data. We take two approach to distribute data to protect privacy; vertical distribution and horizontal distribution. In vertical distribution approach all data values for different attributes resides in different places and in horizontal distribution databases records resides in different places. We use these techniques in combination to improve the classification accuracy as well as persevering privacy in more efficient manner.

II. RELATED WORKS

Secure multi party computation problem comes under where there are different user's want to conduct a joint computation to generate a result without disclosing their data inputs. There computation could be in three possible cases: firstly between the trusted parties, secondly between the partially trusted parties and lastly between the competitors. So in all these situation privacy becomes the main issue. This applies to many areas such as privacy preserving in databases. Privacy preserving in intrusion detection, privacy preserving in statically analysis, privacy preserving in geometric computation and privacy preserving in data mining.

As the data mining is emerging field leads to the knowledge discovery process that is generating useful information (result) from the previously unknown data. This gives rise to the data mining techniques classification, data clustering, mining association rule, data generation summarization and characterization for privacy preserving. The history of secure multi party computation was previously observed by Yao's in two milliner's problem, in which they want to know who is richer without disclosing their wealth. He proposed a protocol which has a secure computation results in one way function. The developed protocol says that Alice has public one way function and the inverse is known to her only and same with Bob, he has public one way function known to him only. Alice and Bob send string to each other one by one. When the string came from Bob, Alice matches it with her own sequence.

In this way a protocol function value is operated privately. The privacy is maintained [2]. Ode goldreich *et. al*. gives a solution to the multiparty protocol problems by a probabilistic algorithm that does not disclose any information about the players with honest players. They used the trusted party for maintaining privacy issues for individual and correctness. There exists a secure solution for any functionality. The approach for operating/ calculation function F follows a combinatorial circuit and then the parties' runs a short protocol for every gate in the circuit. Every participant gets corresponding shares of input wires and the output wires for every gate. The size of the protocol depends on the size of the input. So as it is applicable to the large input likewise in data mining [3]. Lindel and Pinkas examines the work in privacy preserving in data mining instead of using generic protocol, they proposed a protocol for very large database. Their work strategies composed of two large databases then by union of the databases D1 & D2 respectively; perform data mining technique classification by decision tree on ID3 algorithm. For this a database must be having a discrete transaction and a class attribute, row represents the transaction and column represents the attribute.

The ID3 algorithm selects the best predicting attribute by comparing the entropies given as real numbers. They make use of cryptographic tools oblivious transfer, oblivious polynomial evaluation [4]. Same as author Lindel and Pinkas, Rakesh and Srikant previously used classification technique for privacy preserving in data mining. They used the decision tree classifier by two methods; value class membership and value distortion which maintain the privacy as a result in the data distribution from original to the reconstructed distribution differ by measuring the accuracy of classifier. The data perturbation is applied by randomization function [5]. By this entire, privacy becomes the main concern in data mining. There are limited techniques available for this to include gives a classification hierarchy which includes data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. In this entire dimension, privacy can be maintained by their own level. Data distribution dimension can work centralized as well as for the distributed. In distribution cases they can divide into two categories vertical data distribution and horizontal data distribution. There are cases where the data mining will not work efficiently when the voluminous data is centrally located

a. sometimes it is difficult to manage large volume of data centrally.
b. Collection of the data from different sources then combining this data to produce results.
c. Different organization does not want to disclose their data but after applying the data mining at their sites they want to share results.

So perform local computation and combine the intermediate data than to share the produced results. Number of cases exist where the sensitive information lies includes health care organization, collaborative co-operation, multinational co-corporation. As we know there are two cases for privacy preserving in distribution data mining: vertical partitioning and horizontal partitioning. In vertical partitioning the attributes values are placed at different sites. Single entity attribute resides on other places in this way they are vertically partitioned. In horizontal partitioned the transaction are divided on different places. The record lies on different places. Clifton uses the vertical partitioning for privacy preserving in data mining. He makes use of association rule mining, considered mining of Boolean association rule. Transaction can be in {0, 1} '0' shows the absence and '1' shows the presence of transaction/ attribute value. As we have discussed the privacy preserving can be applied to many cases we take medical data as it contains more sensitive information because it is directly related to human personal information. Again this data separation technique for privacy preserving in classification of medical data mining used, horizontal and vertical partitioning. They proposed an algorithm for both vertical and horizontal partitioned data. In vertical partitioned the algorithm achieves privacy but partitioning accuracy is little scarified but in horizontal partition of data, privacy as well as classification accuracy is achieved [6, 7, 8]. Medical domain makes full use of data mining for classification and prediction. The classification technique applied on the LUNG cancer data set to check survival rate. They first prepared the input data set by data pre-processing, data transformation, REFIEF attribute selection, confusion matrix and 10- fold cross-validation on LUNG cancer data and then applied the ADA boost algorithm by checking in the measure of accuracy, sensitivity and specificity. By this, they result in the cancer survival patterns and improved prediction model [9]. For the medical data mining the ethical and legal issues are also of concern because of data keeps the personal information of human. This gives the practitioner and researcher a different scope to use the data mining technique to apply [10, 7].

III. ARCHITECTURE

The proposed architecture gives the description of hybridization partitioning of local and global data separation technique in data mining, fig. 1 depict this. The algorithm works in two phases firstly for local mining and then global mining. In local mining the separation is based on vertical partitioning in which individual data owner perform local data mining to generate local rules. After generating the local rules, these rules are given to the second phase for global mining which uses the horizontal partitioning, in which all data owner gives these local rules to the third party to further classify the decision rules and generate global rules.

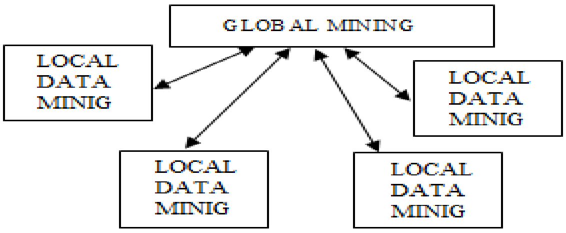

Figure 1. Combining vertical and horizontal Partitioning in Privacy preservation

A. Phase 1: Local mining:

In local mining the individual data owner gives his data set to third party by using the vertical partitioning. In vertical partitioning the table is vertically fragmented in sense of attributes division. Attributes are classified to preserve the privacy or to protect the data as shown in fig. 2.

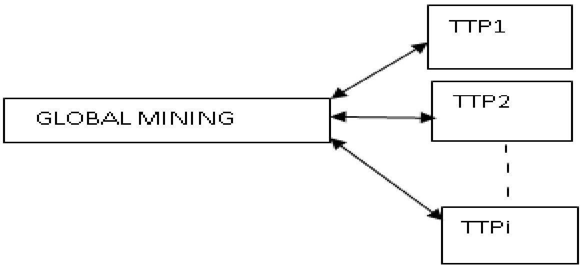

Figure 2. Local Mining (Vertical Partitioning)

Trusted Third party (TTP) applies operation on the partitioned data and gives the result which is in form of rules to the data owner. Data owner analyze the rules. This phase completes the local mining.

B. Phase 2: Global Mining:

In global mining the Different data owners (DO) gives these local rules to third party by using horizontal partitioning. In horizontal partitioning, data owners partition the local rules horizontally which generate a set of rules.

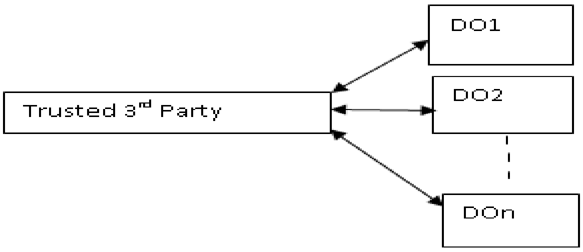

Figure 3. Global Mining (Horizontal Partitioning)

These set of rules are given to third party by different data owners, then third party can only be able to generate the global rules by analyzing the set of local rules without knowing the original data set as shown in fig. 3. This phase preserve the privacy as well as improves the classification accuracy which is little inconsistent in vertical partitioning.

IV. INFORMAL DESCRIPTION ALGORITHM

The medical data set of Lung cancer data from GenBank repository having 699 records with 9 attributes and one class attribute with two possible values: benign and malignant. We are considering only 100 records with all 9 attributes and one class attribute. In this record some values are missing and we used the average of numerical values of those attributes belongs to that class. Input of the algorithm will be medical data sets from r different data sources M1, M2 up to Mr represents the medical data set from data owner 1, data owner 2 up to data owner r respectively and the M1, M2, M3 up to Mn represent m attributes for r data owners. From step 1 to step 3 we perform local mining.

In which, step 1 shows vertical partitioning, in this we generate the n subset with one different attribute removed from Mi at each time where Mi represent the data set for ith data owner and do this for all r data owners.

In step 2 we get m decision trees for each medical data set corresponding each data owner represented by Di={D1,D2,D3,……….Dm }. Output of step 2 will be the input for step 3 which is r sets of decision trees by Ensemble them, find r decision function which will be the local rules for data owners at individual level. Here local mining is completed and global mining taking place.

In step 4, by using the horizontal partitioning these rules are given to trusted third party to generate the global rules in form of trees or decisions and this process ended in step 5. The global rules are distributed to data owners by trusted third party.

V. FORMAL DESCRIPTION OFALGORITHM

/*to classify Medical data set by applying data portioning classification technique*/

O: Data Owner

Oi: ith data owner where i = 1 to n

M: Data set with m attributes

Mi: Represent data set from ith Data owner

Mij: Represent subset of ith data owner by removing jth attribute where j = 1 to m

D: Final Decision Tree or final global rules (Horizontal Partitioning)

Di: Final local rule or decision tree from ith data owner. (Vertical Partitioning)

Dij: Represent decision tree for ith data owner by removing jth attribute.

Input: Medical data set from n different data owners.

Output: Identify more simplify and more accurate decision rules or global rules to classify medical data set.

Begin

/*Local Mining: vertical partitioning to generate local rules*/

Step1: Generate m subset with one different attribute removed from M at each time.

/*Getting Mi = {Mi1,Mi2 ,Mi3 ,………, Mim }*/

Step2: Appling average rule to generate m decision tree corresponding each subset.

/*Getting {Di1, Di2, Di3,………., Dim}*/

Step3: Generalized the decision function Di = {Di1,Di2, Di3,………., Dim}

via majority vote of the m decision trees.

/* Getting {D1, D2 , D3,………, Dn} */

/*Global Mining: Horizontal partitioning to generate global rule*/

Step4: Assemble the final decision function D = {D1,D2, D3,………, Dn},

via majority vote of the n decision trees from step3.

Step5: Classify all n datasets by the final decision function.

End

/*getting more simple and more accurate decision rules to classify medical data*/

VI. PERFORMANCE OF AVERAGE RULE

There are total 699 records with 9 attributes, each attribute having a numerical value. It means we can do any mathematical operation on them. Here, basically we are doing simple average rule: Taking the average of all attribute value for a single person and classify it as benign and malignant.

The Average Formula that we apply here is

$$\text{Average}= \sum_{j=0}^{m} Mij \quad \text{where } i = 1 \text{ to } n$$

However, accuracy of all 699 records is 96.85% which is much better and the way of classification is simpler than other.

VII. CONCLUSIONS

In the field of medicine data mining has become an effective and efficient method to take decisions. The issues of privacy preserving are critical in data mining. The concept can be used to improve accuracy both vertical and horizontal partition. We can protect the privacy with increasing classification accuracy in local as well as global mining. These overcome the accuracy problem in vertical partitioning.

VIII. REFERENCES

[1]. W. Du and M. J. Atallah. “Secure multi-party computation problems and their applications: A review and open problems”, In New Security Paradigms Workshop, Cloudcroft, New Mexico, USA, September 11-13 2001, pp 11–20.

[2]. Andrew C. Yao,”Protocols for Secure Computations”, In Proc. 23rd IEEE Symposium on the Foundation of Computer Science (FOCS), IEEE 1982, pp 160-164.

[3]. O. Goldreich, S. Micali, and A. Wigderson. “How to play any mental game - a completeness theorem for protocols with honest majority”, In proceeding 19th ACM Symposium on the Theory of Computing, 1987, pp 218–229.

[4]. Y. Lindell and B. Pinkas, “Privacy preserving data mining”, In proceeding of Advances in Cryptology – CRYPTO 2000. Springer-Verlag, Aug. 20-24 2000, pp 36–54.

[5]. Rakesh Agrawal and Ramakrishnan Srikant, “Privacypreserving data mining”, In Proceedings of ACM SIGMOD on Management of data,Dallas, TX USA, May 15 - 18 2000, pp 439-450.

[6]. Chris Clifton, “Privacy preserving distributed data mining” In ACM SIGKDD Explorations , November 9, 2001.

[7]. Gang Kou, Yi Peng1, Yong Shi2, and Zhengxin Chen**, “**Data mining of medical data using data separation-based technique” Data Science Journal, volume 6, supplement, 30 July 2007, pp S429-S434.

[8]. Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis “State-of-the-art in Privacy Preserving Data Mining” In the proceeding of SIGMOD Record, Vol. 33, No. 1, March 2004, pp 50-57.

[9]. Jaree Thonakam. Guandona Xu, yachun Zhang, Fuchun Huang, “LUNG cancer Survivability Via Ada Boost Algorithms”, In proceeding of 2nd Australasian Workshop on Health Data and Knowledge Management (HDKM 2008), Wollongong, Australia, pp 55-64.

[10]. John F Roddick, Peter Fule, Warwick J. Graco, “Exploratory Medical Knowledge Discovery: Experiences and Issues” In the proceeding of ACM SIGKDD Explorations Newsletter, Volume 5 Issue 1, July 2003, pp 94-99.

[11]. Asha Khatri, Swati Kabra, Shamsher Singh,”Architecture for Preserving Privacy During Data Mining by Hybridization of Partitioning on Medical Data” 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, DOI 10.1109/AMS.2010.31