Distributed algorithm for privacy preserving data mining based on ID3 and improved secure sum

Ehsan Molaei
Master of IT and data security
Imam Reza University
Mashhad, Iran

Hossein Vadiatizadeh
Bachelor of IT
University of Kerman
Kerman, Iran

Amirmahdi mohammadighavam
Master of IT
Azad University of Kerman
Kerman, Iran

Neda Rajabpour
Master of IT
Azad University of Kerman
Kerman, Iran

Fatemeh ziasistani
Master of IT
University of Gilan
Gilan, Iran

Abstract— **Data mining is a science which is introduced for getting new and useful knowledge from massive data collections. So miming algorithm inputs are always data, but in all cases data have not been stored in one place and in some cases data have been distributed among different servers. For example in some medical data mining we need to cooperation between some hospitals or medical centers, in these cases sharing of data and samples is illegal. According to the HIPAA rule sharing of patient information is unauthorized. So in this paper has focused on privacy preserving of confidential data in data mining process. Suggested algorithm in this paper is a safe distributed algorithm which is using improved secure sum algorithm and performed on classic ID3. By considering algorithms based on tree in data mining, suggested algorithm in this paper have changed a bit and are presentable on all of them.**

Keywords- Privacy preserving data mining, Distributed data mining, Medical data mining, secure sum, ID3

I. INTRODUCTION

Data mining technology has been separated as artificial science by the aim of extracting of correct and new knowledge from massive data [1-3]. This science has been attracted by specialists in various courses and has been greatly grown by considering its vast usage [4, 5].
In classic data mining, data are collected in a server of an office or in research unit of a company or in similar institutes, and then according to needed results, one or more data mining algorithms executed on them. After several years of data mining age and by developing of information and communications technology and understanding of cooperation concept among companies and institutes, a new branch of this science has been introduced as distributed data mining. In distributed data mining unlike to classic data mining, different collections of data have been placed on various servers in order to extract knowledge data mining algorithms should use all data from all distributed data [1, 6].

Using data collections on distributed servers is not as simple as centralized mod, and it is under specific conditions [7]. In order to answer to needs of distributed research, some by increasing of researches and expanding of using data-mining and as a result the development of the use of the distributed data mining, discussion about preserving of the privacy became more important. It has become the most important challenge in this science [2]. So some government and non-government organizations turned against distributed data mining and have made rules in this area. For instance, according to a rule called HIPPA, it is forbidden to publish medical information.

Many papers based on several attitudes in this field have been represented and improved so far which they can be divided into three groups. The first group is the algorithms based on perturbation and transform data, in order to be prevented from the direct reach of the mining server to the main servers. This group of algorithms is considered as the first represented algorithms in the field of privacy, their method of action is in a way that first the main servers make changes in data servers, for example, by adding noise, coding and so on, and then these data are transferred to mining server and at last the process of mining is done on them [9]. The second group is algorithms which are based on making models and each server makes a local model from its own data instead of sending the data or sending transformed data. Then each server sends this model to mining server. The third group is an approach which for preserving of privacy instead of above approaches tries to preserving the privacy with cooperation between servers and distributed the calculations.

The first approach that is the creating of perturbation never could increase the correctness and privacy together and their main complexity has been the creating balance between these two factors [2, 10]. The second algorithms are those which their basis is creating of local models, they preserve the privacy well, but they could never guarantee the accuracy, they have their own limitations too, for instance

never have presented an algorithm for integrating local models. Although the represented algorithms in this paper are based on the third attitude, we have tried to publish no extra information. In this paper by improving of a method called secure sum, and using it by a secure protocol for communications between data and mining servers, the distributed algorithm has been represented based on (ID3) algorithm. In this way in section 2 we talk about several cases about related works in this area and then in section 3 first we describe some definitions that are necessary and then in section 4 we introduce the secure sum briefly and then its improvement on it that we presented. In section 5 we introduce the ID3 algorithm briefly and afterwards in that section we represent the suggested distributed algorithm in section 6. At the end of paper we present the theoretical evaluating of the suggested method in section 7 and we compare our algorithm with same algorithms.

II. Related work

Many algorithms based on various techniques in the field of privacy preserving data mining have been discussed, but by considering page limitations of the essay, we will only mention some of basic methods and similar methods related to our work. In methods of data perturbation, as one of these works, we can refer to the reference of [11] which in this essay has used the randomize data perturbation and then data perturbation has been made on these data. In the other paper, Mukherjee et. al. tried to show by combination of data perturbation and parameters transformation, they were able to represent a method which is resistant against the attack recovery against each of two techniques [13], a problem which exists in all methods based on perturbation is the contrast between two problems of accuracy and privacy preserving since the methods that used strong perturbation which generally make data different, and reduce the accuracy of data mining [24, 25, 26, 10].

In another article by Yu et. al, in order to achieve privacy preserving, the SVM has been used. In this essay, each server has made a local model based on SVM and in order to mine these models they should be shared their models and finally with combining these models they make the final model based on all available properties in all servers and it is used for mining [14]. Although this method preserves the privacy well, according to the ambiguous aspects in combination step, this approach was not appreciated.

In other attitudes, an attempt has been made in suggestion of distributed solutions for the problem of privacy preserving data mining instead of concentration of centralized solutions [15-17]. So that in these methods instead of attempts in data perturbation or attempts for making local models and their transfer instead of main data they try to distribute calculations. So that each local server only works on its data and at last, results in different servers are combined together. These approaches don't have the ambiguities of previous approaches. In an essay that proposed based on this approach shanek et.al introduced a method to find the nearest neighbor in a distributed and private manner and employed it to implement several algorithms of private mining. K-nearest neighbor is one of these algorithms. The problem raised in their work is disclosure of new sample which is in contrast with the privacy principle in these works [15].

In another paper in this field, we repeatedly refer to [16], a distributed algorithm for mining association rules. In this paper they suggested that distributed mining servers by using the safe and privacy algorithms make association rules for each needed parameters and finally by combining these parameters, we can get association rules [16].

III. Defenition And Background

In this section we define several concepts that they are necessary to understand the algorithm and its environment:

A. Data partitioning methods

The way that data partitioned is one of most important factor in distributed data mining. In fact all algorithms are designed based on the type of data partitioning. Generally, there are two types of data partitioning, vertical partitioning and horizontal partitioning. 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 we want to collect different data such as financial, medical, insurance and housing data about different people resident in a city. In this process we 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 [2, 18, 19].

Another type of partitioning called horizontal partitioning the data are partitioned so that the same set of data about different entities are distributed over different places. For example suppose that in a data mining project we want to investigate the effects of a drug on patients having a special disease and in order to increase the number of our samples we need to obtain the same information about this issue from different medical centers. In such settings it is said that the data are partitioned horizontally. In this paper we will work on horizontally partitioned data [18, 2, and 19] in this essays, we work on Horizontal partitioning data.

B. Work environments

In the distributed data mining there are two types of working structure. The first is s2s or server to server model,

in this type of settings all cooperating parties are in the same level and in order to perform data mining is no need to coordinate with a higher authority and they cooperate only with each other. Another model is call c2s or client to server settings in which the cooperating members are not in the same rank and the members send the data to the mining servers following their request to perform data mining. In this paper the proposed algorithm can be implemented compatible with both type of define model [20, 21]. In this essay, the suggested algorithms are adaptable to both above defined environments.

C. Security foundations

By considering the distributed feature of the suggested method in this essay, it is important to ensure from identity of the received information resource. In this paper for insurance of servers identity, we will use the digital signature based on the elliptic curve problem which nikooghadam et al, in the reference of [22] has introduced as mentioned in this paper the used methods have been suitable both from point makes time-saving in our algorithms. In order to protect the transferred information symmetric encoding AES is proposed in this paper which can significantly enhance the security measures and also is suitable in terms of runtime. The asymmetric encoding RSA can be used to publication the symmetric key.

IV. Introduction Of Secure Sum And Improving On It

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, we can have a reference to [19, 14] 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 we can refer to the collusion of two servers for accessing to the information of server that is between them. In this paper, we 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. For example we calculated the sum of three numbers x, y and z without disclosure none of them is. The method of secure sum does this work by adding a random number like R. The method works in a way that the first server adds its number x with a random number R and sends it to the second server, since the sent number to the second server is the sum of two numbers x and R ,this server is not able to recover the private number x. Now the second server adds its number y to the sent number and sends it to the next server, Combining of results continues to the last server and finally, the last server after adding its number to the received number sends results for the first server. In this step the first server receives a number which is the sum of all numbers and the random number R which has been produced by the first server itself. Now, first server must subtract the random number R from final summation that received to it and obtain summation of all numbers without random number R. In figure 1, we have an example for secure sum algorithm for better understanding [23].

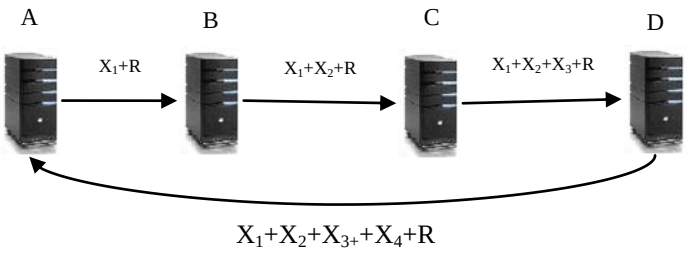

Figure 1.Schema of secure sum algorithm

The above algorithm has some disadvantage and advantage but one of its disadvantages is that in a system by collusion between two servers, they can find number value of server that is between these two servers. For example in above figure if B and D want to find x_3, B can send x_1+R to D and D will get $x_1+x_2+x_3+R$ from C, then D can subtract number that get from B and C and find x_3. To avoid this problem we improve secure sum and proposed an algorithm. Our algorithm prevents collusion between two servers and transforms collusion between two servers problem to collusion between three servers. Figure 2 show improved secure sum algorithm work structure.

The above algorithm has some disadvantage and advantage but one of its disadvantages is that in a system by collusion between two servers, they can find number value of server that is between these two servers. For example in above figure if B and D want to find x_3, B can send x_1+R to D and D will get $x_1+x_2+x_3+R$ from C, then D can subtract number that get from B and C and find x_3. To avoid this problem we improve secure sum and proposed an algorithm. Our algorithm prevents collusion between two servers and transforms collusion between two servers problem to collusion between three servers. Figure 2 show improved secure sum algorithm work structure.

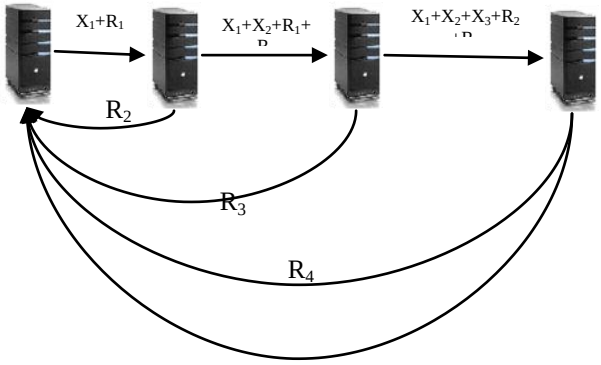

Figure 2: schema of improved secure sum algorithm

V. A BRIEF REVIEW ON ALGORITHM ID3

A. Abbreviations and Acronyms

Algorithm ID3 is one of the most practical algorithms of data-mining which is based on the construction of decision tree for classification of new samples. This algorithm has started to calculate its essential parameters by using available samples and after calculating of these parameters in each step, it completes one level of decision tree. Essential parameters in this algorithm are entropy and gain, which this algorithm by calculating them in each level acquires the best field for continuing of the tree construction. In the following we describe work structure of this algorithm.

1) the construction of the tree from top to down and at first all the samples have been located in S.

2) entropy of S is calculated with formula 1.

$$\text{entropy}(S) = \sum_{i=1}^{i=c} -Pi \log_2 Pi$$

3) in the third step for each remained feature that it is not chosen, the gain parameter is calculated by formula 2.

$$gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv)$$

4) after calculating gain for all remained features, the feature which has got the most gain is selected as a root candidate of the present level, and the S collection for each branch is formed based on selected feature.

5) the above operations continue by the time that the tree completes. Completed tree is dependent on one of below conditions.

a) All available samples in one node belong to one class.

b) There are no features for new division.

c) There is no sample.

As you see, the above algorithm in each step needs the calculation of two parameters. Gain(S,A) and entropy(s) which by calculating and using them it tries to make the best classifier tree.

VI. THE SUGGESTED DISTRIBUTED ALGORITHM

Suggested solutions in this paper have represented the ID3 distributed algorithm in order to solve the privacy preserving problem. To calculate parameters such as entropy and gain, instead of gathering data on a server, this approach will calculate parameters by getting help from local data servers. The suggested algorithm, by using mechanisms such as secure sum, digital signature and data encryption will guarantee the privacy preserving. In the following we will describe the work structure of our proposed algorithm.

A. The first phase, key distribution

Since there are only two kinds of relationships in our system which has been shown in figure 3, there is a direct relationship between the mining server and each of the data servers, in order to create security in communications, we will use the symmetric encryption to providing the security and high performance.

So, AES cryptography algorithm which is known as a standard algorithm is suggested in this paper and in order to change the symmetric key between each data server and mining server RSA asymmetric cryptography algorithm is suggested. By considering the purpose of this paper which is the preserving of data privacy and not representing of a secure protocol to communicate between servers, we won't concentrate on the details of this step and only describe this step briefly.

The distribution phase and setting the keys is done before starting of mining process and all the agreements will be fixed to the end. The symmetric keys between the mining server and each of the data servers are called SK_{msi} that i is index of each data server. The symmetric key between two data servers called SK_{DSij}, mention that SK_{DSij} and SK_{DSji} are the same. In Figure 3 an encrypted communication is shown.

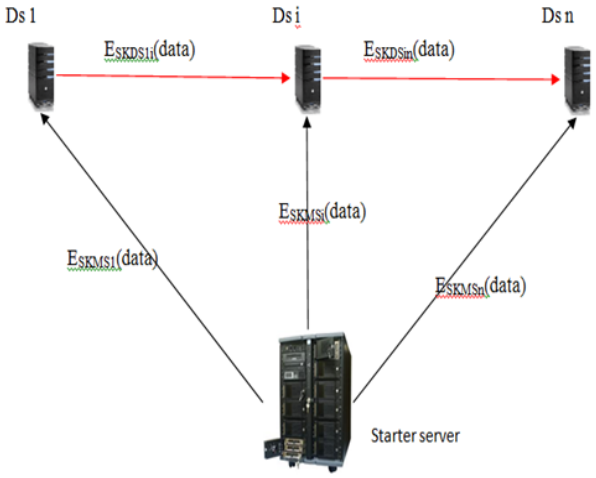

Figure 3 An encrypted communication in system

B. The second phase: Making features the same

First, in this step, mining in order to make available data bases the same has been distributed on data servers, applying general features of each data base such as sample numbers, feature numbers and the name of each property signed and encrypted by symmetric keys SK_{msi} are sent to each data servers . The mining server sends the mining features list to data servers which Signed and encrypted by the key of SK_{msi}.

C. The third phase: construction of the Tree

As you observed in previous part and in explanations related to ID3 algorithm, for construction of the tree we need to calculate, gain, for each remained feature in each level of tree, so it should be in a way that it can calculating this parameter distributed. By paying attention to the gain calculation formula, which has been mentioned below, we see that for each feature like A we need to calculate gain with formula 3.

$$gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv)$$

And also we need to calculate several parameters separately which include the entropy collection of all samples, the entropy of each amounts for feature A, and the other numbers of total samples and the total number of whole samples which amount of A features is equal to v.

To calculate each of these amounts, we will use the improved secure sum which has been explained in part 4. By considering the mining environment which is called S2C here, the algorithm of the secure sum starts by mining server, to better understand, in figure 4 the way of operation has been determined, algorithm calculate each of essential parameters as summation of local parameters. For example algorithm calculates total instances count with formula 4.

$$|s| = \sum_{i=1}^{i=n} instance\ count(i)$$

So all needed parameters are sum of local values which they can be combined with suggested secure sum and we can acquire the total sum. It is clearly obvious that mining servers by having the total sum of each parameter can cover the decision tree of all available samples on all servers. We showed one round of improved secure sum algorithm in figure 4.

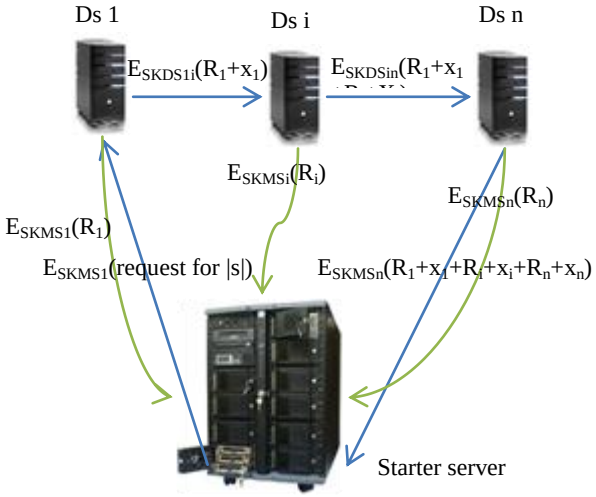

Figure 4.one round of improved secure sum algorithm

D. The fourth phase: constructing of final model

In this step the decision tree has been completed and the final model for data classification has been constructed. Now if the mining server has got a new sample, it can use the available tree for classification. This tree can be used by other servers. Since none of these available data on data servers have been appeared and on the other word the decision tree won't appear the excess information except the needed information for classification.

VII. THEORETICAL EVALUATING OF THE SUGGESTED METHOD

Privacy preserving data mining are investigate with three attitudes which are security and privacy evaluation, evaluation of time complexity and the evaluation of network communications. In this paper our suggested algorithm is based on the classic ID3 algorithm which has eager learning nature, so construction of tree model which is the most complex part of this algorithm is done only one time and later servers that having this tree model are able to classify new samples by using it.

A. Performance and communication evaluation:

The ID3 algorithm is a greedy algorithm that tries to select the best choice for data division in each step, the best implementation for this algorithm is recursive implementation. Supposing that the feature numbers in our data are m and total samples on all servers are n ,and most numbers of valued amounts for our features are t, the most depth of constructed decision tree will be m. On the other hand by considering the recursive nature of algorithm, in each level, algorithm needs t call for each values of selected feature. And By solving this formula we can find time step

which is equal to $O(t^m*n)$, although it seems exponential, it is reasonable and applicable because data sets dimensions are limited.

$$W(m) = t*W(m-1) + O(n)$$

The suggested algorithm needs great communications for construction of tree model by considering the nature of distribution and using the module of the improved secure sum which we can divide these communications in each request to communication between data servers and p-2 communications between data servers and mining server. P used as numbers of data servers. These communications between servers in comparison to secure sum, which needs p-1 communication for communication between the data servers and 2 for communication between data servers and mining server is more than, but according to our arrangement in this suggested algorithm which is the better privacy preserving and security, these communication numbers are acceptable.

B. The privacy evaluation

Our purpose in this paper was proposed the algorithm that has least information disclosure and in this way distributed calculations approach has been used. The algorithms which have used our same approach for classification, have got the problem of disclosure of the new samples, in other words instead of using the secure sum method, we have used another method based on it without the probability of collusion [15]. In this algorithm by considering not changing of the main data unlike to methods based on perturbation and gathering data on mining servers which always encounter to the recovery risk, the main data will always be preserved. The table 1 has a brief comparison between the suggested algorithm and the previous methods from point of privacy.

Disadvantages / Algorithm	The imbalance between accuracy and privacy	Recovery and Disclosure data	Disclosure New instance
Data perturbation approach	weak	weak	strong
Modeling approach	weak	strong	strong
Calculation distributed approach	strong	strong	weak
Proposed algorithm	strong	strong	strong

TABLE 1: a brief comparison between proposed algorithm and previous methods

VIII. 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. In this paper, our approach by using the attitude of distributed calculations on local data servers instead of gathering data on a central server have represented the algorithm which in group of classification data algorithms can compete with other algorithms which provide privacy. In this method unlike the previous algorithm correctness and privacy are preserved the parallel and perfectly and it covers the great weakness of many algorithms based on distribution of calculation, that is the appearance of the new sample.

In future, we intend to represent an algorithm which is shared between the distributed data mining and centralized data mining and is compatible with famous algorithms of traditional data mining.

REFERENCES

1. F. Giannotti, D. Pedreschi, Mobility, Data Mining and Privacy, springer, book, 2007
2. C.C. Aggarwal, P.S. Yu, Privacy-Preserving Data Mining-Models and Algorithms, springer, book, 2008
3. S. Dua, X. Du, Data Mining and Machine Learning in Cybersecurity, CRC press, book, 2011

4. S. Mukherjee, M. Banerjee, Z. Chen, A. Gangopadhyay, A privacy preserving technique for distance-based classification with worst case privacy guarantees, Elsevier, Data & Knowledge Engineering 66 (2008) 264–288
5. V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, Y. Theodoridis, State-of-the-art in Privacy Preserving Data Mining, ACM,2004
6. S. Dua, X. Du, Mining and Machine Learning in Cybersecurity, book, Auerbach Publications Taylor & Francis Group, 2011
7. B.C.M. FUNG, K. WANG, R. CHEN, P.S. YU, Privacy-Preserving Data Publishing: A Survey of Recent Developments, ACM Computing Surveys, Vol. 42, No. 4, Article 14, 2010
8. R.P. Winn, Confidentiality in Cyberspace: The HIPAA Privacy Rules and the Common Law, 33 Rutgers L.J. 617 (2001-2002)
9. C. Keke, L. Ling, Geometric data perturbation for privacy preserving outsourced data mining, springer, Knowledge and Information Systems December 2011, Volume 29, Issue 3, pp 657-695
10. C. Keke, L. Ling, Privacy Preserving Data Classification with Rotation Perturbation, IEEE, Data Mining, Fifth IEEE International Conference on, 27-30 Nov. 2005,
11. R. Agrawal, R. Srikant, Privacy-preserving data mining, ACM SIGMOD, pp 439–450, 2000
12. A. Deutsch, Y. Papakonstantinou, Privacy in database publishing, ICDT, pp 230–245, 2005
13. K. Liu, H. Kargupta, J. Ryan, Random projection-based multiplicative perturbation for privacy preserving distributed data mining, IEEE Trans Knowl Data Eng 18(1):92–106, 2006
14. H.Yu, J.Vaidya, X.Jiang, Privacy-Preserving SVM Classification on Vertically Partitioned Data, springer, Volume 3918/2006, 647-656, 2006
15. M. Shaneck, Y. Kim, V. Kumar, Privacy Preserving Nearest Neighbor Search, springer, Machine Learning in Cyber Trust, pp 247-276, 2009
16. M. Kantarcioglu, C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 9, SEPTEMBER 2004
17. X. Yi, Y. Zhang, Privacy-preserving naive Bayes classification on distributed data via semi-trusted mixers, Elsevier, Information Systems 34 371–380, 2009
18. E.Magkos, M.Maragoudakis, V.Chrissikopoulos, S.Gritzalis, Accurate and large-scale privacy-preserving data mining using the election paradigm, ELSEVIER, Data & Knowledge Engineering 68 ,1224–1236, 2009
19. M.Kantarcıoglu, J.Vaidya, Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data, ICDM Workshop on privacy preserving data mining, 2003
20. A.C. Aggarwal, P.S. Yu, Privacy-Preserving Data Mining Models and Algorithms, springer, book, Springer Science+Business Media, 2008
    B. C. M.FUNG, K.WANG, R.CHEN, PH. S. YU, Privacy-Preserving Data Publishing: A Survey of Recent Developments, ACM, ACM Computing Surveys, Vol. 42, No. 4, Article 14, 2010
21. M. Nikooghadam, M.R. Bonyadi, E. malekian, A. Zakerolhosseini, A protocol for digital signature based on the elliptic curve discrete logarithm problem, journal of applied sciences 8(10):1919-1925,2008
22. C. Clifton, M. kantarcioglu, J. vaidya, X. lin, M.Y. zhu, Tools for privacy preserving distributed data mining, ACM SIGKDD Explorations Newsletter, Volume 4 Issue 2, December 2002
23. H. Luo, J. Fan, X. Lin, A. Zhou, E. Bertino, A distributed approach to enabling privacy-preserving model-based classifier training, springer, Knowl Inf Syst, 20:157–185, 2009
24. H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, privacy preserving properties of random data perturbation techniques, IEEE ICDM, 2003
25. Z. Huang, W. Du, B. Chen, Deriving private information from randomized data, ACM SIGMOD, 2005