A Novel Algorithm for Privacy Preserving Distributed Data Mining

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

Mehrdad Jalali
PHD of computer engineering
Azad university of Mashhad

Hossein Vadiatizadeh
Bachelor of IT
University of Kerman

ABSTRACT

With the development of data mining science and information technology, distributed data mining was considered. Distributed data mining was submitted with different purposes, such as increasing the accuracy of result and using data from multiple sources. With the development of distributed data mining, challenges in this field were introduced soon. The main challenge was the issue of privacy preserving. In the past few years, solutions for this problem have been proposed but each one has had a weakness.

Given the importance of knowledge, today's organizations need to Distributed Data Mining, Our goal in this article is to provide an approach that able to preserve the privacy in distributed mining. Our approach can be implemented on most of the algorithms. Proposed approach has been useed data encryption and some of techniques that used in security.

General Terms

Privacy preserving distributed data mining algorithm.

Keywords

Privacy preserving, data mining, cryptography, ID3, Bayesian classification

1. INTRODUCTION

Nowadays that the volume of data and data generation devices are increasing every day, knowledge Known as the most important wealth in the organizations and institutions. Traditional organizations had huge archives of paper that uses of them are very difficult. But this does not mean that today's digital archives are used and mined well. Many organizations have data that alone are not useful, or with sharing data, organization can extract newer and more knowledge. Just distributed data mining can extract knowledge from distributed and shared data in these conditions [1, 2].

Distributed data mining as well as any other science, is faced with challenges. The most important challenge in this field is privacy preserving. In this issue many laws have been enacted, for example one of the most important rules named HIPPA. According to this law, any disclosure of personal information about patients is crime and health institutions are responsible for the personal information of patients [3, 4, and 5].

Given the importance of data mining in development of other sciences, and leading decision makers in organizations and institutions, today, organizations are often required in data mining and its new field called distributed data mining[. Hence science researchers in data mining provide solutions to the problems of distributed data mining. One of these problems was to preserve the privacy of data that Participate in mining process. To solve this problem solutions offered include data perturbation and randomization, modeling, distributed calculation and etc. but each solution has weaknesses. The first approach methods are using data perturbation; this approach has serious problems about privacy preserving and increasing accuracy parallel that these problems have never been solved. The second approach is modeling. This approach has never provided a comprehensive method for integrating models. Third approach based on distributed calculation. Although this approach partially solves the problems of previous approaches, but many algorithms of this approach needs to disclosure of new instances that it is a weakness in privacy preserving concept [4,6,7,8].

In this paper, algorithm have been tried to preserve the privacy with use of security tools. The main idea in this method is that transformed data in data servers with cryptography module. After this process this algorithm uses some other security tools for providing privacy preserving data mining algorithm.

2. RELATED WORKS

With the development of data mining and provide the various methods for privacy preserving, [8- 16] Advantages and disadvantages of these methods and how to implement them better has been much discussed. Most of the proposed methods based on perturbation, randomization or anonymity. [4] The main disadvantage of these methods was loss of accuracy in return for increased privacy.

Paper [14] proposed modeling method along with SVM for privacy preserving. In this method data server constructs their local model and send and share it for mining, and construct global model for prediction with these local models. Major disadvantage of this approach is that it didn't provided global method for integrating local models.

In the third approach, researcher proposed distributed calculation to solve the privacy preserving problem. In papers [12, 15, 16] be seen some methods with this approach that how do these algorithms is such that try distribute the calculations between data servers and finally get the results of calculation instead of data. Major disadvantage of this approach is that many of these algorithms have to disclosure new instances.

This paper has been proposed an approach that tries to solve privacy preserving in data mining by use of some security tools. For example this algorithm uses encryption module instead of perturbation and transformation module and it uses trust server for publication encryption keys. This paper has been tried to propose an approach that can implement that on each classic data mining algorithm.

3. BACKGROUND AND DEFINITIONS

3.1 Data Partitioning Methods

The way that data partitioned is one of most important factor in distributed data mining and in fact 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 goal is collecting different data such as financial, medical, insurance and housing data about different people resident in a city. In 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 goal is investigating the effects of a drug on patients having a special disease and in order to increase the number of samples will be required to obtain the same information about this issue from different medical centers. In such settings it is said that the data are partitioned horizontally [2, 10, and 13]. This paper has been worked on horizontally partitioned data.

3.2 Working Settings

In the distributed data mining there are two kind of working structure. The first is s2s or server to server model, in this type of settings all collaborate parties are in the same level and in order to perform data mining is no need to adapt with a higher authority and they adapt only with each other. Another model is call c2s or client to server settings in which the collaborating members are not in the same level 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 [4, 9].

4. PROPOSED IDEA

Dataset have contains different fields and different data types such as numeric data, classification data, nominal or other data types. Numeric data which different samples such as binary data, numeric data discrete and continuous are classified, In between some of existing data that data mining algorithms work with repeat them, Not with perform mathematical calculations such as distance and etc, Can be encoded or encrypted somehow that in other servers cannot be detected and retrieved. Exploring operations must be designed somehow that final server by using sequence search, compare fields and the number of iterations ability to retrieve data or even ability prove the existence a particular sample is not in the dataset.

For nominal data, discrete numerical data and binary and in general for all categorical data, distance calculation function can be represented as follows:

$$\text{Distance}(a,b) = \begin{cases} 0, & \text{if } a=b \\ 1, & \text{otherwise} \end{cases}$$

This means that, if a and b are similar, the distance between them are zero and if they are dissimilar they have maximum distance. So algorithms can use encryption instead of data perturbation. In this way, each server encrypted its data with unique key and sends to third-party server. Since the distance between original data with together and equivalent encrypted them with together not difference. Processing results is identical on both [17].

In the next section, it has been explained how to implement proposed idea along with examples on two famous algorithm classifications ID3 and Bayesian.

4.1 Implementation on ID3 Algorithm

In this algorithm, our system will have mining server, data servers, CA server which task generates and send encryption keys and decryption. requester Server may be one of data servers. Overarching theme system is shown in Figure 1.

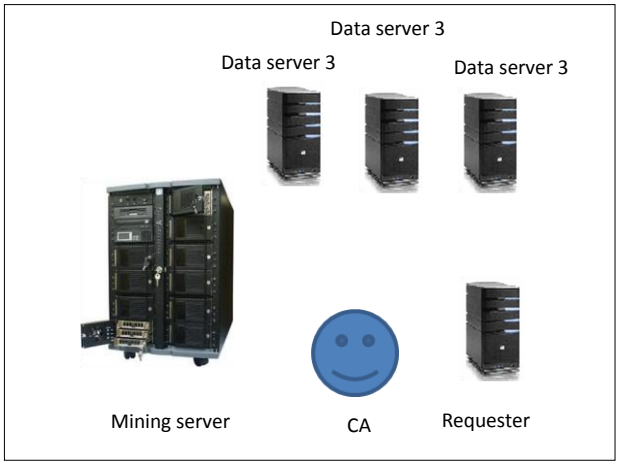

Fig 1: Scheme of system

The algorithm is carried out in several steps, in first step demantant sent signed and encrypted message to the CA. This message contains a demandant ID and mining request and also ID of each data server. In the second step CA server generated key encrypt and decrypt but only sends encryption key to the data servers. In the third step, data servers, encrypted their data with encryption key, and sent to the mining server. In the fourth step, the mining server receives the encrypted data and by using these data creates ID3. In the fifth step, ID3 decision tree as a tree classification sent to demandant server and finally, demantant server for classification new instances, encrypt its new samples by using encryption key and by using Decision Trees received, obtain class of new instances. Steps of Implementation this algorithm is shown above in Figure 2.

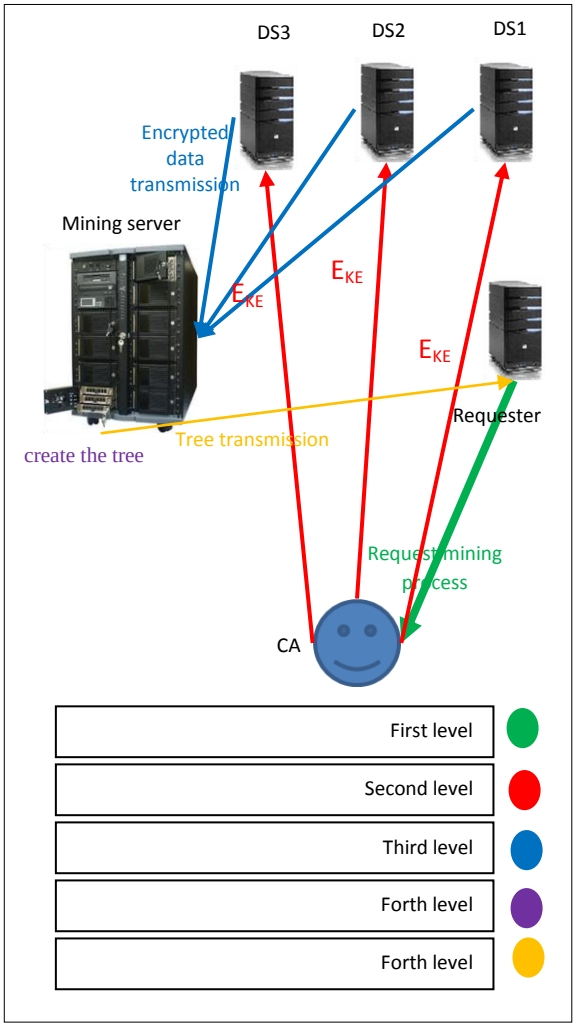

Figure 2: Implementation steps of algorithm based on ID3 classic algorithm

According to the description above and Figure 2, is not need to disclose decryption key. Because the demandant server can using encryption key for encrypts new sample and then classification it.

4.2 Implementation on Bayesian Algorithm

In this section like previous section, system has mining server, data server, CA server which task generates and send encryption and decryption keys to other server. Figure 1 is an overview of the system. Implementation steps of proposed method on Bayesian algorithm is as follows:

In the first step, demandant server request message containing its ID and also data servers ID to be signed and encrypted, send for server CA. In the second step CA server generated encrypt key and decrypt key but only sends encryption key to the data servers and demandant server. Then in the third step, demandant server encrypt new sample, and also data servers encrypt their sample. Note that all servers with a shared encryption keys are doing encryption operation and encrypt similar data are the same. After encrypted the data, this data is sent to the demandant server, that in fourth step, obtain encrypted class of new sample By using encrypted data received from the data server and this class for decrypted sent to CA server. In the fifth step, class Label is decryption in CA and is sent for the demandant server. Different steps of algorithm are shown in Figure 3.As shown in the figure 3, CA server must be trust.

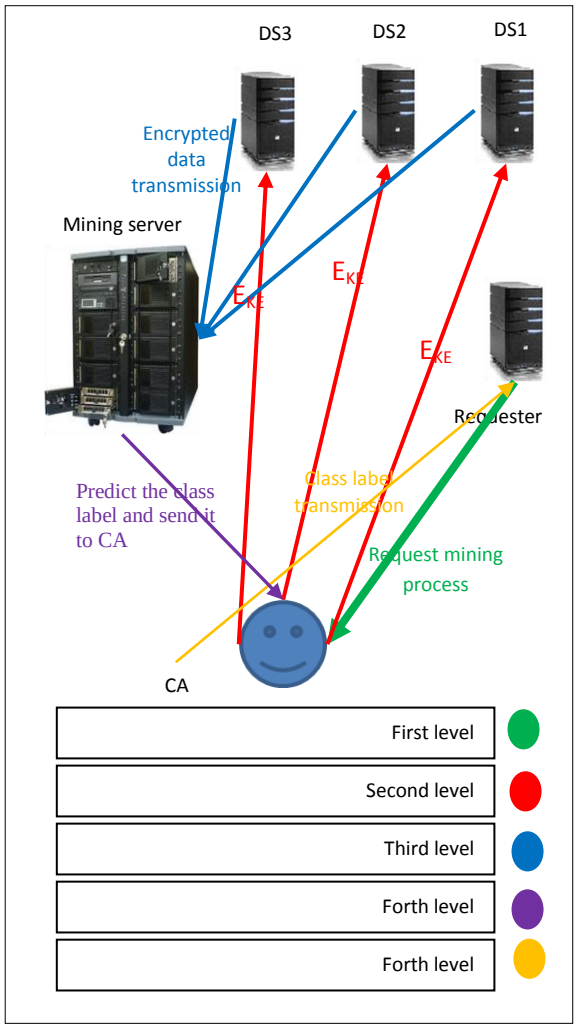

Fig 3: Implementation steps of algorithm based on Bayesian classic algorithm

5. EVALUATION AND PROOF OF PERFORMANCE

The proposed algorithm consists of two main phases, the First phase, all security measures such as data encryption and their displacement and second phase, perform data mining by using one of classical algorithms, here for instance is used from Bayesian and decision tree algorithms. Terms of the complexity time there are several various Cryptography, that software implementation some of them is optimal, although can be used coding instead of encryption and complexity time improved as much as possible. However, the second phase for implement depends on the choice of the classical algorithm explored, that in here for instance is given an eager algorithm and a lazy algorithm.

In distributed data mining algorithms as also previously have brought as purpose of this paper, maintain confidentiality is the most important principle and since that encryption only once can be executed for every query explore, and the other hand for example in ID3 algorithm after construction tree, can be used repeatedly for classification, global costs of our approach is acceptable. For declare weakness and power of our approach it has been provided a simulation and

prototyping based on ID3 classification algorithm. Simulation has been used the dataset in table 1 for simulation.

Table 1. Dataset used in this example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

With the execute ID3 algorithm on a dataset of table 1, will be construct the ID3 decision tree that shown in figure 4.

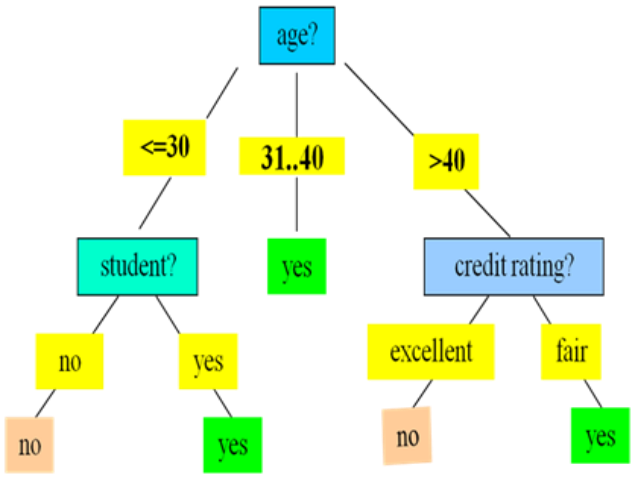

Fig 4- ID3 tree

To facilitate the implementation of the algorithm it has useed below coding instead of encryption:

Age=f1, income=f2, student=f3, credit-rating=f4

<=30 = a1, 31...40 = a2, >40 = a3

High=i1, medium=i2, low=i3

No=s1, yes=s2

Fair=c1, excellent=c2

With above coding, dataset in table 1 transform into table 2 that shown below.

Table 2. Part of data set in table 1 that it has been encoded

f1	f2	f3	f4	Buys-computer
a1	i1	s1	c1	no
a1	i1	s1	c2	no
a2	i1	s1	c1	yes
a3	i2	s1	c1	yes

In the next step with execution ID3 algorithm on coded dataset, it will be constructed the coded ID3 tree that shown in figure 5.

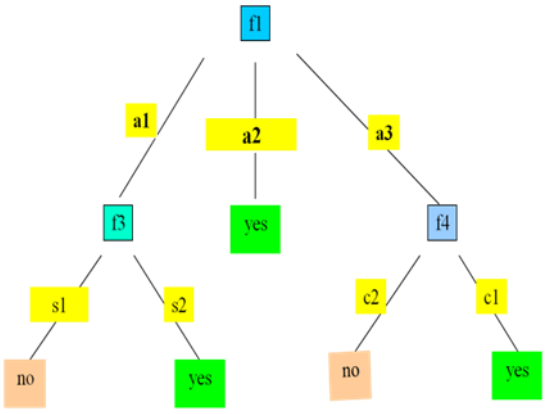

Fig 5: Encoded tree on encoded data set in table 2

It can use these instructions for Bayesian algorithm. As seen trust CA server in this approach has high importance and as disadvantage can cite to be collusion between the server CA and the mining server. It should be emphasized that many security protocols have trust CA server that is acceptable and reduces the cost of security and other complexities.

6. CONCLUSIONS

In many applications information technology, maintain confidentiality and privacy preserving as a challenge always posed and have been proposed solutions for it. From applications where privacy has become to a challenge is distributed data mining, that in this paper, It has been proposed an algorithm that use coding and encryption instead of data perturbation, that provide lower costs and full accuracy, and also it can use based on all classic algorithm.Although this paper has been tried to fix the biggest weakness of algorithms that are based on the data transformation that they can't increase the accuracy and privacy parallel together.

7. REFERENCES

[1] Dua, S., Du, X., Data Mining and Machine Learning in Cybersecurity, CRC press, 2011

[2] Jiawei, H.n, Micheline K., data mining: concepts and techniques, second edition, Elsevier, 2006

[3] Lindell, Y., Pinkas, B., " privacy preserving data mining", journal of cryptology, springer, 2002

[4] Aggarwal C., C., Yu, P. S., Privacy-Preserving Data Mining-Models and Algorithms, springer, 2008

[5] Winn, P,A., Confidentiality in Cyberspace: The HIPAA Privacy Rules and the Common Law, 33 Rutgers L.J. 617 (2001-2002)

[6] Giannotti, F., Pedreschi, D., Mobility Data Mining and Privacy, springer, 2007

[7] Evfimievski, A., Gehrke, J., Srikant, R., "Limiting Privacy Breaches in Privacy Preserving DataMining", ACM sigmod record, 2003

[8] Brankovic, L., Islam, Md.Z., Giggins, H., "privacy preserving data mining", security, privacy, and trust in modern data management, springer, 2007

[9] FUNG, B. C. M., WANG, K., CHEN,R., YU,PH. S., " Privacy-Preserving Data Publishing: A Survey of Recent Developments", ACM, ACM Computing Surveys, Vol. 42, No. 4, Article 14, 2010

[10] Magkos, E., Maragoudakis, M., Chrissikopoulos, V., Gritzalis, S.," Accurate and large-scale privacy-preserving data mining using the election paradigm",ELSEVIER, Data & Knowledge Engineering 68 ,1224–1236, 2009

[11] Mukherjee, S., Banerjee, M., Chen,Zh., A.Gangopadhyay, " A privacy preserving technique for distance-based classification with worst case privacy guarantees", Data & Knowledge Engineering 66, 264–288, 2008

[12] Yi, X., Zhang, Y., "Privacy preserving Naive Bayes classificationon distributed data via semi-trusted mixers", Elsevier, Information Systems34, 371–380, 2009

[13] Kantarcıoglu, M., Vaidya, J., "Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data", ICDM Workshop on privacy preserving data mining, 2003

[14] Yu, H., Vaidya, J., Jiang, X., "Privacy-Preserving SVM Classification on Vertically Partitioned Data", springer, Volume 3918/2006, 647-656, 2006

[15] Kantarcioglu, M., Clifton, C., Privacy-preserving distributed mining of association rules on horizontally partitioned data, IEEE Transanctions on Knowledge and Data Engineering 16 (9) (2004) 1026–1037

[16] Shaneck, M., Kim, Y., Kumar, Vipin, " Privacy Preserving Nearest Neighbor Search" springer, 2009, Machine Learning in Cyber Trust, pp 247-276

[17] Inan, A., Kaya, S.V., Saygin, Y., Savas, E., Hintoglu, A. A., Levi, A., privacy preserving clustering on horizontally partitioned data, Elsevier, 2007, Data & Knowledge Engineering 63 (2007) 646–666