A Novel Method for Preserving Privacy in Big-Data Mining

Nasrin Irshad Hussain
Kaziranga University
Jorhat, Assam
India

Bharadwaj Choudhury
Kaziranga University
Jorhat, Assam
India

Sandip Rakshit
Kaziranga University
Jorhat, Assam
India

ABSTRACT

As our daily lives has become digitization which led to an explosion in the collection of data by organization and individuals. Organization push their vast amount of data into big data clusters, but most have implemented zero security measures. Protection on confidentiality of these data is very important. In recent years privacy-preserving data mining has been emerged as a popular research area for the security of sensitive information in the network. In this field we study the extraction of knowledge or pattern from big data maintaining the commercial or legislative privacy constraints. Privacy preserving mining of distributed data has diverse applications. We have numerous algorithmic techniques for privacy preserving data mining.This paper presents a privacy preserving method for big data. We proposed a novel method for preserving privacy in mining big databases. Our goal is to mining the data while preserving data privacy and confidentiality.

Keywords

Privacy-preserving data mining,security, distributed data, big data

1. INTRODUCTION

The proliferation of information technology and the internet in the last few years take it easy to find anything or any individual's information. The term big data usually describe the exponential growth of data for both structured and unstructured. And is important to business and society as the internet has become. With the rapid advancement in the technology, it will create large databases, carrying large amount of information. Production of data is increase so rapidly such that world's volume of data doubles every eight-teen months for example, an enterprise data are predicted to increase by about 650% over the next few years [1, 2]. To extract knowledge from big data, there are many data mining techniques. A new era of research started where existing data mining techniques are considered for privacy preserving. As the development and use of internet increases the threat against privacy and it create serious problem.

The organizations or the companies which have their big data about the ongoing activities of their clients. To make use of the data, the data owners utilized data mining techniques to extract knowledge and compromise the privacy of their client [4]. It is very challengeable to mining new knowledge while protecting individuals' privacy. If data owners wish to release the data mining output but assured that they are not giving any identity of their clients.

The privacy preserving data mining have two different considerations [3]: (i) modification of raw data such as id, password, name, address, age etc. in order for receiver not to compromise any person's privacy. (ii) By using data mining techniques, mining the sensitive knowledge from a databases without compromising the data privacy.

In this paper, we propose a cryptographic algorithm for preserving privacy of raw data. We are in a data rich situation, if these data are not analyzed or mined to gain knowledge then it will have no use. But while mining the knowledge it is very important to preserve privacy of customers. So it is desired to build a model of data without violating the privacy of individual records, e.g., compute average age before knowing the age of any one person. We use cryptographic approach to big data for privacy preserving data mining. In our approach we preserve the privacy of raw data and after mining, only desired output is to be published. The primary goal is to design a method which is to be effective without compromising security.

2. DATA MINING ON BIG DATA

Data mining or knowledge discovery of interesting patterns from large collections of data, has been considered as an important area of database research. Big data are not just a huge volume of data in terabytes. Other important characteristic in addition to volume, include variety, velocity and value [5]. Big data is changing security analytics by providing new tools and opportunities for large quantities of structured and unstructured data [6]. Big data include unstructured data such as text, audio, video and website log files etc. follow real-time streams for analysis to maximizing the business value by making the decision to real-time. If we are not gaining knowledge from big data, means lost data, knowledge as well as money. So it will beneficial for companies to mining their big data stores without violating the clients' privacy.

Data mining is a process of extracting knowledge or pattern from big data. The process of mining big data is now viewed as a threat to the privacy preserving. In data mining process, the algorithms require all participant in a distributed system to follow a privacy model and make assumption on behavior of participating node. But it will not possible for big data. In this paper, we developed a novel framework for privacy preserving data mining on big data, where all participants are divided into some category based on some conditions and after that will go for mining.

3. BIG DATA ANALYTICS FOR SECURITY

New big data applications are becoming part of security management software because they can help to clean, prepare and query data in heterogeneous, incomplete and noise formats efficiently. Analysis of big data provides companies the power to identify trends and improve business. Big data analytics changes their landscape to improve information security and situation awareness. As the era of big data begin, data plays an important role, so we have to give security for it. Big data analytics can analyse financial transactions, log files, and network traffic to identify anomalies [7]. Fraud detection is one of the most visible use for big data analytics, so there is a requirement for system architecture which give security to big data. Security might not have been as necessary big data clusters were accessed only by some small groups of programmers, but for wider enterprises or organizations it is difficult to share all data with all levels of employee. Big data framework and tools are now commoditizing the deployment of large-scale, reliable clusters and therefore are enabling new opportunities to process and analyze data.

4. HOW SECURITY IS IMPORTANT FOR BIG DATA

As IT emerges, the size of databases increases too fast, so it is difficult to handle such a huge amount of data. The big data includes structured, unstructured and semi-structured data. In this paper we discuss security prospect of big data. Now we are in data rich situation where data plays an important role and consider as “Gold” [14]. So security is very important to these data from unauthorized used. For any business, they take care of their customers and their personal details which if reveal it will create problems for its’ survival. Privacy of individuals is to be preserved within the organization. Data mining or other techniques are used on big data, but it will be taken care of that these will not been effect any individual.

5. RETATED WORK

In this section, we report some of the relevant works on privacy preserving scheme.

5.1.1 *Agrawal and Srikant's scheme* [8] considered a decision tree classifier from training data in which the values of individual records have been perturbed by adding random values from probability distribution. After this the data records look very different from original records and distribution of data values also look very different from original. Then there is a problem to accurately estimate the original values in individual’s data records, for this problem they proposed a novel reconstruction procedure to accurately estimate the distribution of original data values with some loss of information. But the authors say that this is acceptable for practical situation.

5.1.2 Oliveira and Zaine [9] considered some geometric data transformation to study the feasibility of achieving PPC. They revealed that basic transformation is feasible only after normalization of data because data transform through this methods would change similarity between data points. So clustering of data is useless. Distortion methods adopted to successfully balance privacy and security in statical databases are limited when the perturbed attributes are considered as a vector in the n-dimensional space.

5.1.3 Inan, Saygin, et al's scheme [10] ensures accuracy based on the dissimilarity matrix construction using a secure comparison protocol for numerical, alphanumerical and categorical data. Here the communication cost is high because of the involvement of the third party.

5.1.4 *Teng and Du* [11] gives an approach which takes advantage of the strength of both SMC (Secure Multi-party Computation) and randomisation approaches to balance the accuracy and efficiency constraints. They implemented method for ID3 decision tree algorithm and association rule mining problem. This approach achieve a better accuracy compared to the only randomization approach and more efficient than the SMC approach.

5.1.5 *Secure Multi –party Computation* [12] based on clustering vertically partition data. In vertically partitioning data, the attributes are split across the partitions. This work ensures the privacy while limiting communication cost.

5.1.6 *Kalita, Bhattacharyya et al's scheme* [16] used three transformations- translation, rotation and reflection successfully in combination. The authors established a secure and accurate scheme after applying the hybrid perturbation technique. In this technique, reflection based transformation is helpful to improving the intruders’ complexity significantly.

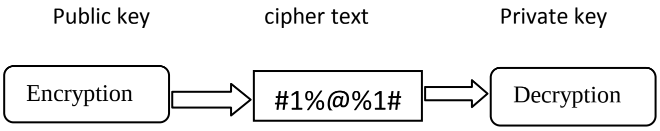
There are several novel efforts for solving the problem of privacy preserving data mining. The main goal is to provide security to the data.

6. BASICS OF OUR APPROACH

6.1 Cryptography

In PPDM, data mining provides objectives, cryptography provides tools. To protecting information by transforming it or encrypting it into an unreadable formate called cipher text. Only those employee who have a secret key can decipher or decrypt the message into plain text. As the electronic communication become more prevalent, electronic security become increasingly important. Cryptography is used to protect e-mail messages, credit card information and corporate data.

Cryptography system can be classified into two broad category-symmetric-key system and public-key system. In symmetric-key system, one single key is used by both sender and receiver. And in public-key system, a public key is known to everyone and private key that only the recipient of messages uses.



We used asymmetric cryptography for encrypting the customers information which have greate importance in business. Asymmetric or public-key cryptography is a class of cryptographic algorithms,which require two different key-one is secret or private and other is public key. The public key is used to encrypt the paintext and private key is used to decrypt the cipher text to plain text. The public key may be published without compromising security but private key used only by the authorized people. In comparison to symmetric cryptography, it is much more secure process. We used digital signature which include public and private key. It is most popularly used in asymmetric cryptographic because it give high security to the data. After you have the public-private key, using public key the data set will be encrypted and the authorized people can decrypt it for data mining purposed. Otherwise the data privacy is maintained.

6.2 Partitioning based on clustering

In our approach, we used Rule system for clustering the data. Rule system is efficient and capable of working with large volumes of data. And it is tolerant to noise and work with numeric as well as nominal attributes. Based on some rule, it will classifies the data into some cluster to mining that data.

Example:-

Table1: Sample table for Insurance Company

Name	**Address**	**Age**	**Annual income**	**Disease**	**Insurance**
Ramesh	Bombay	28	1,50,000	No	No
Mukesh	Guwahati	36	3,00,000	Heart problem	Yes
Shikha	Pune	42	1,80,000	Physically handicap	Yes
Abdul	Delhi	30	5,00,000	No	No
Samir	Pune	40	4,60,000	Diabetes	Yes
Jacky	Bombay	32	1,20,000	Diabetes	No
Nisha	Delhi	37	8,00,000	Skin problem	Yes

We first analyse the data, and looking for condition that will enable us to determine whether the insurance should be yes. To do this, select the attribute that covers the higher percentage of instances that meet the desired result i.e. people take insurance. Then selection process give some rules based on which clusters are formed. If an attribute has many values, then rule systems will permit non-exhaustive coverage.

6.3 Digital Signature

Signature is generally used to authenticate documents. Digital signature are used to authenticate the electronic documents. It ensures that the original content of the document that has been sent is unchanged. Digital signature is used any kind of message, whether it is encrypted or not, so that the receiver can be sure of sender's identity and the message arrived intact. To sign a document, you must have a digital ID which is obtained from various certification authorities on the web. Digital signature is a small block of data that is attached to your document [18]. It is digital ID which include the private and public key, the private key is used to apply the signature on the document. The signature verifies that the document has not been altered since it was signed.

7. PROPOSED METHOD

In our approach, we give a data security platform secures data and access control for your data. To safeguard the data, encryption and key management is required. Allowing only approved access to the data by the users. Only authorized user who have the private key can see the plaintextThe proposed method first classifies the raw data-set into some clusters based on the rules. For clustering, we used Rule system. For different data-set, there are different rules based on some condition viewed in the data-set. There are no any rigid rule which have to be followed by all data-set, so it is efficient for large scale data.In our Accessment method, there are three layer for accessing the data. Data privacy is the main concern, so the data in every layer have to be secure and some authorized employee can access the data. Everyone in the organization have one single public key and every layer have one private key based on which they can used or see the data.

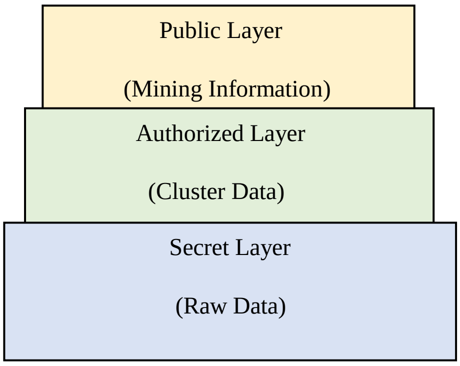


Fig1: Levels of Accessment Model

In the **secret layer,** mainly encryption is done. After customer give his/her details to the company then it will be encrypted and stored in the database. So that no one can retrieve any ones' private data,such as name, address, phone number, credit card details etc. In this layer all incoming cutomers details are saved for future data mining process. Encrypting the data is done by using RSA asymmetric crytoalgorithm. Here using public key the data is to be encrypted. After this, the database administrator can put digital signatures on the customers information data, which is unique and is very difficult to forge.

In the **authorized layer**, the employee which have the private key first decrypt the digital signature and varifies the sender's identity. After this, they can decrypt only needed customer details for data mining process. Data are first analysed and then a data mining technique is used. Here we used clustering for mining the data. Clustering is done by some rule system. After analysing the data,we got some rules,based on which we can create clusters. Usually data mining technique is used on large data for mining unknown and unexpected information. Clustering is very efficient work on big data mining. Taking the above example to illustrate working process of the layers. Authorized employee or data-miner taking only the data which are to be analysed for mining. Then, they have the table of content listed below.

Table2: Analysed Table for data mining:

Age	Annual income	Disease	Insurance
28	1,50,000	No	No
36	3,00,000	Heart problem	Yes
42	1,80,000	Physically handicap	Yes
30	5,00,000	No	No
40	4,60,000	Diabetes	Yes
32	1,20,000	Diabetes	No
37	8,00,000	Skin problem	Yes

In this layer, the selection process will work, select conditions which enable us to determine

If ? then insurance = yes

Selection process:

Age<35	0/3
Age>= 35	4/4
Income< lakh	1/3
Income>2lakh	3/4
Disease=suffer	4/5
Disease=No	0/2

If Age>=35 then insurance = yes

We now try to properly classify the other instances covered by the rule

If Age>=35 and ? then insurance = yes

If Age>=35 and income>2lakh then insurance = yes

We now try to properly classify the other remaining instances covered by the rule

If Age>=35 and income >2lakh and Disease=suffer then insurance=yes

Now we have a rule that classifies perfectly the three instances it covers. Based on this clusters are formed.

In the **public layer**,the mining information or decision which have been taken after data mining technique is viewed. This layer is accessable from all the employees within the company or organization. The privacy of data is maintain throughout the procedure. Only authorized people can see the individual's information. And also the data mining technique is used on clusters without violating the privacy of customers.

8. PROCESS DETAILS

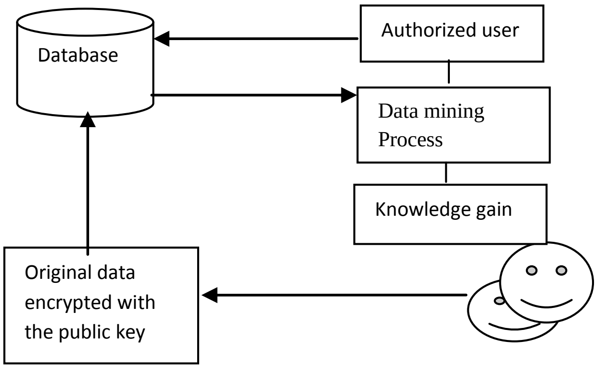


Fig2: Data Mining Process

Stepwise Process:

Step1. Apply for digital ID which include private and public key by the database administrator.

Step2. Customer information will be encrypted using public key before save to the database.

Step3. The administrator apply the digital signature on the customer document so that no one can alter any information of the customer when it is send to the data miner.

Step4. When authorized employee, he/she can decrypt the signature and verify the identity. After that data has to be decrypted and apply data mining technique.

Step5. The mining knowledge or pattern is published to the public layer.

Description of the method:

- **Privacy protection on raw data-**

  In the first step, we encrypt the customers' data for safeguard data from unauthorized use and preserving privacy. For encryption of the data, we used RSA algorithm and using public key encrypt the data. After encryption, data are safely stored in the database. Except authorized employee, all people can see the encrypted data.

- **Data Mining process-**

  When organization willing to apply data mining on their big data for gaining knowledge, then authorized employee of the organization, take required data from database for mining. Using private key they can only decrypt the required raw data. After decryption, a data mining technique call clustering is used. Clustering is efficient for big data mining. It will put all data into some clusters and extract some knowledge from it. For clustering we used rule system to build the clusters based on these rules. From the above example the authorized employee have got the analyzed table for data mining without violating privacy policy.

- **Knowledge Gain-**
  After data mining, knowledge discovery process is ended. We have gain some patterns for uplifting your business, so that it will be published in the Public layer of the accessment model. Here all employee can see the mining knowledge and according to this, they will work. Preserving privacy is the main aim of our work. If organizations or companies follow this type of model then it will be very beneficial for them and they can assure their customers' that their details are very important for them and they got full privacy against this. Customers' data are the wealth of a company, so preserving privacy for these data is very important.

9. CONCLUSION

Now in business environment, having access to the right information means making the right decision critical to surviving. Business need to protect their information as it accumulates much faster because of big data [12]. For any business their customer and their information are asset, so it have to be secure. As the era of big data begin, companies databases also increases so it is not possible to take one by one data and analyzed it based on some rules made. Therefore we used data mining technique on big data to extract knowledge. Data mining process work on data and data contain information about individuals [18].

10. REFERENCES

[1] Gartner, Post event brief, Gartner IT Infrastructure, Operations and Management Summit 2009, Orlando, FL. available at www.gartner.com June 23–25 2009

[2] IDC, Digital data to double every 18 months, worldwide marketplace model and forecast, Framingham, MA. available at www.idc.com May 2009.

[3] data mining using Matrix Algebraic Approach",doi:10.4156/jcit.vol4.issue3.5.

[4] Arie Friedman,”Privacy preserving data mining”pp.4,January 2011

[5] P. Russom, Big Data Analytics, Best Practices Report, Fourth Quarter, The DataWarehouse Institute , Renton, WA, September 18 2011.

[6] Big Data Analytics for Security Intelligence,september 2013

[7] R. Agrawal, R. Srikant, “Privacy-preserving data mining”, In: Proceedings of the 2000ACM-SIGMOD on management of data, Dallas, TX, USA, May 15-18, 2000

[8] S.R.M. Oliveira, O.R. Zaiane, “Privacy Preserving Clustering By Data Transformation”, In Proc. Of the 18th Brazilian Symposium on Databases, Manaus, Brazil, October 2003, pages 304-318.

[9] A. Inan, Y. Saygin, E. Savas, A. Hintoglu, A. Levi.: Privacy-Preserving Clustering on Horizontally Partitioned Data. Data Engineering Workshops, 2006.

[10] Z. Teng, W. Du, “A hybrid multi-group approach for privacy preserving decision tree building”, In: Proceedings of the 11th Paci_c-Asia conference on knowledge discovery and data mining (PAKDD2007).

[11] B. Pinkas, “Cryptographic techniques for privacy-preserving data mining”, SIGKDD Explore, 2002, 4(2): 12-19

[12] “Big security for big data”, available at www8.hp.com/ww/en/secure/pdf/4aa4-4051enw.pdf

[13] N.I.Hussain,P.Saikia,B.Choudhury, S.Rakshit,”Study of a Decision tree approach to analyse Big data”,pp.2,published on Micro-2014.

[14] Brian Lent, Arun Swami, Jennifer Widom,”Clustering Association Rule”.

[15] Abraham Otero,”Data mining techniques”,pp25-36/39.

[16] M. Kalita, D.K. Bhattacharyya, M. Dutta, “Privacy Preserving Clustering - A Hybrid Approach”, In: Proceedings of the ADCOM’08, Chennai, December 2008.

[17] Alaa H Al Hamami, Suhad Abu Shehab, “ An Approach for Preserving Privacy and knowledge In Data Mining Apllication”, Journals of Emerging trends in computing and information science,vol. 4,No1 jan 2013.

[18] Searchsecurity.techtarget.com/digital-signature