# RSA Based Secret Key Generation for Privacy Preserving Association Rule Hiding

Salvath Rumi[1], Dr. Shreedhara K S[2]
[1]*Computer Science and Engineering, UBDTCE, Davanagere - 577001 Karnataka, India*
[2]*Professor, Computer Science and Engineering, UBDTCE, Davanagere-577001 Karnataka, India*

***Abstract*- There is a probability that during the event of data mining, the information can be exposes to several parties and this will lead to the breach of individual privacy. However, this problem can be solved by applying the Privacy Preserving Data Mining. This paper involves two stages: At first, Association rules are generated for the input database using Apriori algorithm. In association rule mining, leakage of sensitive data can cause potential threats to privacy and protection. So to hide the sensitive rules, in the second stage the two techniques that is anonymization and cryptosystem are applied to maintain privacy. Anonymization is a method of protecting the private data of individual using methods like suppression and generalization. Through cryptosystem secret key is generated which encrypt the private data. The type of encryption that is used here is RSA encryption. The result of the proposed system is compared with existing technique that is Particle swarm optimization. The result will then be measured by specific parameters measurements, which are the privacy level and mining quality (data loss).**

**Index terms- PRIVACY PRESERVING DATA MINING, ASSOCIATION RULE HIDIND, PSO**

## I. INTRODUCTION

Nowadays finance department, marketing retails, industries of telecommunication provides lots of details. To look at this data people lack proper time.
The powerful data mining technique is therefore introduced for the data extraction from this huge collection of data. It automatically analyses the data, classifies the data and summaries the data into useful information. The process of retrieving of information [1] utilizes the mining of association rules. The main goal of data mining is therefore to extract the unknown information which includes the number of credit cards, secured information, numbers regarding personal identification, cell phone or telecommunication numbers etc. Privacy Preserving Data Mining (PPDM) with association rule hiding method here which represents the data mining technique saves the sensitive and essential information from the unlawful disclosure of data.
The sanitized database is offered by PPDM along with various essential conditions such as, i) it is not supposed to expose the rules regarding sensitive information and, ii) non sensitive information should only present in the sanitized database and iii) database should not be modified. The quality of data in sanitized dataset will be same as of original dataset. Association rule hiding technique includes approaches such as Meta heuristic, exact approaches, the approaches based on reconstruction, the approaches based on cryptography, border approaches etc. The data which is encrypted is now sent to the opponent party. To the second group or party the data is not revealed in the process of data mining and therefore data privacy is kept secured. In this paper section II shows the background work, section III shows proposed architecture, section IV shows Implementation, section V result and section VI shows conclusion.

## II. BACKGROUND WORK

Pathak et al. (2012) proposed privacy preserving association rule mining using impact factor concept [2]. The impact of factor of a transaction is the number of itemsets that are presents in sensitive association rules. This method did not provide full privacy as it modified fewer transactions.
Yi et al. (2015) proposed privacy preserving association rule mining in cloud computing [3]. In order to overcome the high computation cost of k-anonymity, k-support and k-privacy techniques, user encrypts its data and store it in the cloud and mining

of association rules are done using outsourced semi-honest servers. Compromisation of all servers is tackled by selecting servers from cloud servers.

Modi and Patil (2016) presented privacy preserving association rule mining on horizontally partitioned database with the involvement of trusted third party [4]. This approach securely extracts association rules even the communication channel is unsecure between the parties. It uses elliptic curve based Differ-Hellman and digital signature algorithms to ensure privacy and security.

J. Sumithra Devi and M. Ramakrishnan (2018) used FDM technique to find frequent itemsets. The support count is encrypted using RSA algorithm and forwarded the other sites. One data initiator, one data combiner and other parties as client in ARM process are used. Experimental results show that this method is flexible to some extent but ensures privacy only during global support count calculation process.

## III. PROPOSED SYSTEM

It has been mentioned in figure 4.1.1 the suggested block diagram for proposed privacy preserving association rule hiding method [1].The main goal of the project is to produce the sanitized dataset for the input dataset. This process of converting original dataset to sanitized dataset is called Data Sanitization. Sanitization process removes or encrypts sensitive information in the dataset.

The first step in proposed system is generating the association rules for the input dataset. The Associations rules at the same time highlight the relationship probability between various items inside the data sets which are larger in size. The Apriori Algorithm is used to generate these rules and also to find the frequent set of items this algorithm is utilized.

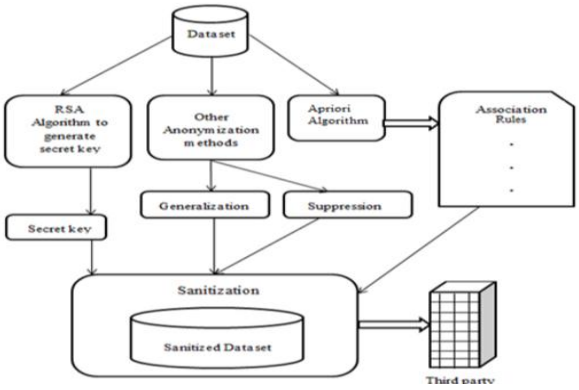


Figure 1: System Architecture

Second step involves utilization of both Cryptographic method and Anonymization methods which is used for preserving the privacy of the sensitive items in dataset. Cryptography algorithm used here is RSA algorithm that generate secret key.

Using that secret key the private data is encrypted. With the secret key, RSA transforms the original database into sanitized database and thereby provide PPDM. In conjunction with RSA other data anonymization methods like suppression and generalization which also preserves the private data are also utilized. Therefore, data base when it is received by a third party is kept save and secured and loss of information in this case is negligible.

## IV. IMPLEMENTATION

This project is developed using Django as a web framework, pycharm as IDE and python as a programming language because as per the rules of syntax that python uses it offers users the concepts of expression and in this case there is no need to use any additional code. The python scores more over other languages used for programming since it is large and has robust standard library. Due to this standard library one can also select the modules of wide ranges as per the requirements.

1. Association Rule Mining

The working of his technique is: Let I = {i1, i2, ... , im } be a set of items, D = {t1, t2 , ... , tn} be a transactions sets in such a way that the proper subset is ti for I. An association rule is represented as X => Y, X is subset of I, Y is a subset of I, and X∩Y=Ø. Support for X =>Y, is signified by S(X=>Y), and this is defined as: Support(X=>Y) = |∩Y| / |D|

Confidence for X =>Y, is represented as:

Confidence(X=>Y) = |X∩Y| / |X|

Apriori Algorithm: The association rule mining is the first phase and for this purpose the generation of rules makes use of Apriori algorithm [13]. One of the most used and well-functioning algorithms is the Apriori algorithm which is used for the mining of association rules. This algorithm works by finding the frequency of individual items in the data base it works and stretch them slowly as these items is seen to appear often in the data base sufficiently. The association rule can be determined by the frequent item sets.

2. Cryptography

Cryptography is all about the protocol construction and analysis which do not allow third parties or other people to get access to the messages which are private and hence the data is kept confidential.

RSA: The algorithm is also the asymmetric cryptographic in nature. It generally means that on two different keys this algorithm works. Those are private key and public key. Procedure:

- The public key is sent by the client to the server and some data is requested.
- The data is encrypted by the server making use of public key of the client and data which is encrypted is sent.
- The data is received by the client who then decrypts this data.

1. Anonymization

The process involves the removal of information which is personally identifiable from the sets of data.

It is a process whereby the hidden or personal information is altered irreversibly in such a way that no identification of data subject is possible either by direct or indirect means and in this way the data controller alone or in combination with other parties cannot identify the data. There are two common methods which are used here: Suppression, in which the attributes values are replaced by asterisk (*) and Generalization, in which the attribute values are seen to be replaced by the category which is broader one.

## V. RESULT

First thing done in execution is uploading the dataset shown in figure 2. Next two prime numbers are entered for generation of public and private keys pairs through which dataset is encrypted. After giving two prime numbers the dataset is encrypted through public key. Encrypted file can be viewed by downloading it as shown in figure 3.


Figure 2: Uploading dataset and giving two prime numbers.

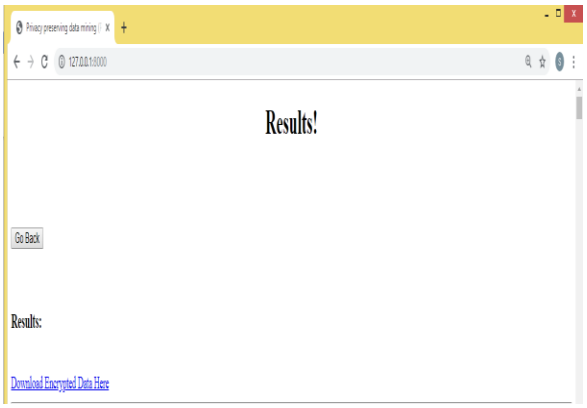


Figure 3: Link to download encrypted file.

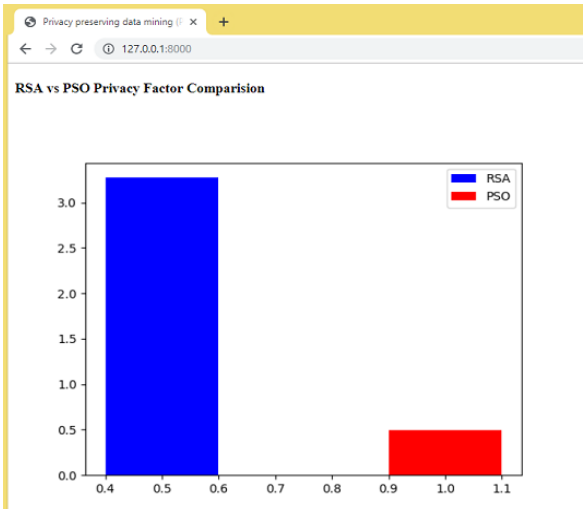


Figure 4: Comparison graph

Figure 4 shows the comparison graph between RSA based privacy preserving association rule mining and PSO.

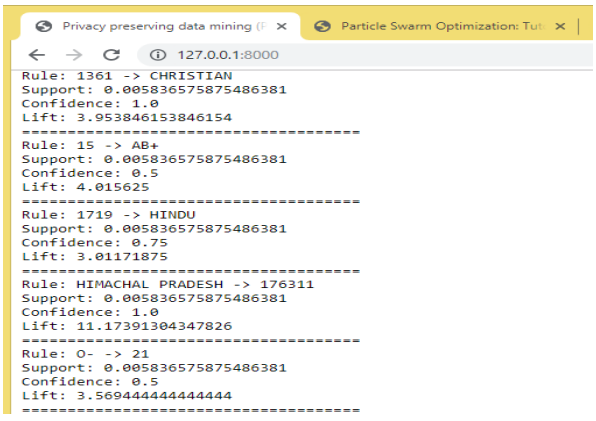


Figure 5: Association Rules

Fig.5 shows association rules generated from dataset. By generating secret key the sensitive rules are kept hidden.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | NAME | AGE | GENDER | SSN | BLOOD GR | RELIGION | SALARY | DISEASES | STATE | PINCODE |
| 2 | AREEZ | 45 | F | 1561 | A+ | CHRISTIAN | 219924 | ARTHRITIS | PUNJAB | 142050 |
| 3 | AIDEN-VE | 46 | F | 1674 | A- | MUSLIM | 209420 | HYPERTEN | PUNJAB | 142050 |
| 4 | CONLLY | 81 | M | 1177 | A+ | MUSLIM | 216834 | HEPATITIS | ANDHRA F | 524203 |
| 5 | GUSTAV | 50 | F | 1195 | AB- | CHRISTIAN | 219340 | CANCER | KARNATA | 591156 |
| 6 | COBAN | 54 | F | 1497 | AB- | SIKH | 205230 | AIDS | NAGALAN | 797116 |
| 7 | ELSHAN | 71 | F | 1732 | AB- | CHRISTIAN | 205462 | STROKE | WEST BEN | 721659 |
| 8 | SZYMON | 50 | M | 1557 | AB+ | SIKH | 217479 | MALARIA | ODISHA | 761114 |
| 9 | JOHANN | 44 | M | 1619 | B+ | HINDU | 212250 | ASTHMA | KERALA | 689597 |
| 10 | DECLYAN | 71 | M | 1323 | O- | CHRISTIAN | 213426 | HEART AT | ANDHRA F | 532462 |
| 11 | ZAYD | 42 | F | 1205 | AB+ | MUSLIM | 205874 | CANCER | BIHAR | 803116 |
| 12 | GHYLL | 93 | M | 1987 | AB+ | CHRISTIAN | 217460 | INFLUENZ | HARYANA | 132102 |
| 13 | JAIDEN | 24 | M | 1106 | B- | MUSLIM | 217667 | DIARRHEA | BIHAR | 845101 |
| 14 | D'ARCY | 70 | F | 1014 | O+ | SIKH | 218747 | SMALL PO | UTTAR PR | 262901 |
| 15 | SALMAN | 50 | F | 1514 | O+ | MUSLIM | 202031 | ARTHRITIS | PUNJAB | 144805 |

Figure 6: Patient Dataset

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | NAME | AGE | GENDER | SSN | BLOOD GR | RELIGION | SALARY | DISEASES | STATE | PINCODE |
| 2 | * | <50 | F | * | A+ | INDIAN | 219*** | 2.85E+21 | PUNJAB | 142*** |
| 3 | * | <50 | F | * | A- | INDIAN | 209*** | 1.83E+30 | PUNJAB | 142*** |
| 4 | * | >=50 | M | * | A+ | INDIAN | 216*** | 1.83E+19 | ANDHRA F | 524*** |
| 5 | * | >=50 | F | * | AB- | INDIAN | 219*** | 2.65E+17 | KARNATA | 591*** |
| 6 | * | >=50 | F | * | AB- | INDIAN | 205*** | 2.85E+11 | NAGALAN | 797*** |
| 7 | * | >=50 | F | * | AB- | INDIAN | 205*** | 1.39E+15 | WEST BEN | 721*** |
| 8 | * | >=50 | M | * | AB+ | INDIAN | 217*** | 4.43E+18 | ODISHA | 761*** |
| 9 | * | <50 | M | * | B+ | INDIAN | 212*** | 2.85E+13 | KERALA | 689*** |
| 10 | * | >=50 | M | * | O- | INDIAN | 213*** | 1.83E+28 | ANDHRA F | 532*** |
| 11 | * | <50 | F | * | AB+ | INDIAN | 205*** | 2.65E+17 | BIHAR | 803*** |
| 12 | * | >=50 | M | * | AB+ | INDIAN | 217*** | 1.56E+22 | HARYANA | 132*** |
| 13 | * | <50 | M | * | B- | INDIAN | 217*** | 2.17E+22 | BIHAR | 845*** |
| 14 | * | >=50 | F | * | O+ | INDIAN | 218*** | 1.39E+20 | UTTAR PR | 262*** |
| 15 | * | >=50 | F | * | O+ | INDIAN | 202*** | 2.85E+21 | PUNJAB | 144*** |
| 16 | * | <50 | F | * | AB- | INDIAN | 230*** | 2.3E+15 | HIMACHA | 173*** |

Figure 7: Privacy Preserved Dataset

## VI. CONCLUSION

The major motive of this project is to discuss a privacy preservation technique that is meant to protect information during the data mining event. The proposed methodology which is seen to work in two main steps and they are named association rule mining and for sanitization the generation of a secret key. The suggested techniques make use of the Apriori algorithm at first whereby the association rules from the data base are mined. RSA algorithm is used in the second phase and this allows the generation of secret key which can convert the data base by this secret key into the sanitized data base so that it will hide the sensitive items in association's rules. Second phase also involves utilization anonymization methods such as suppression and generalization which also preserves the privacy of the dataset. So here the second phase is combination of cryptography and anonymization. After that, sanitized dataset generates the association rules in which privacy is preserved. Performance of proposed system is compared with existing technique that is Particle Swarm Optimization (PSO). Analysis which is given here clearly mentions that the techniques which are suggested here have a lot of efficiency for the data set because no much modification is made to the sanitized dataset as it is much similar to the original database. After dataset is converted to sanitize dataset no information is lost from the dataset.

## REFERENCES

[1] Least lion optimisation algorithm (LLOA) based secret key generation for privacy preserving association rule hiding. D. Menaga1, S. Revathi11B.S. Abdur Rahman Crescent University, Vandalur, Chennai 600048, India

[2] Pathak k, N S Choudri and A. Tiwari, 2012. Privacy preserving assocition rule mining by introducing concept of impact factor. Proceedings of the 7th IEEE conference on Industrial Electronics and Applications.

[3] Yi, X, F.Y Rao, E. Bertino and A Bouguettaya, 2015. Privacy Preserving association rule mining in cloud computing. Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, April 14, 2015.

[4] Modi, C.N, and A.R Patil, 2016. Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases without involving Trusted Third Party (TTP).