International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Secure Rule Mining in Horizontally Partitioned Database

Priyanka Pandey[1] & Javed Akhtar Khan[2]
[1,2]Department of Computer Science and Engineering, Takshshila Institute of Engineering and Technology, Jabalpur, M.P., India
pandeypriyanka906@gmail.com; javedaktarkhan@takshshila.org

Abstract*—*The new era of information communication and technology (ICT), everyone wants to share/ store their Data/ information in online media, like in cloud database, mobile database, grid database, drives etc. when the data is store in online media there is main problem is arises related to data is privacy because different types of hacker, attacker or crackers wants to disclose their private information as publically. For securing that information from those kinds of unauthorized people we proposed and implement of one the technique based on the multilevel security concept with taking the iris database on weka tool. And these papers provide the high privacy in distributed database environments.

Keywords— Data Mining; Distributed Database; Privacy preserving protocols; Association Rule Mining.

1. Introduction

In recent years, the data mining [1] [2] [3] became a very interesting topic for the researcher due to its vast use in modern technology of computer science but due to its vast use it faces some serious challenges regarding data privacy and data privacy became an interesting topic. Many methods techniques and algorithms are already defined and presented for privacy preserving data mining. These privacy preserving techniques can be classified mainly in two approaches Data modification approach [15] [16] [17] and Secure Multi-party Computation approach [5] [6]. Data Mining in last few decades has become very useful as the database are increasing day by day many people now connected with the computers so it becomes necessary for computer researchers to make the data so fast to access, also need to find right data. The term Data Mining emphasize on the fact of extracting the knowledge from large amount of data, So data mining is the process through which we collect knowledgeable data from very large data. Now a days the database are very large which consists so much information but what we want to find is the relevant data from large database or want to find some patterns which becomes very difficult with normal DBMS but with the use of data mining techniques we can find the hidden patterns and information from large database system. So we can also term data mining as the knowledge mining, pattern extraction etc. But before applying data mining techniques we need to apply some processes which we known as preprocessing of data. Although data mining is one of the step involved in process of knowledge discovery but still it becomes more popular by name then that.

2. Proposed Work

Due to the increased demand for knowledge discovery in all industrial domains, it is necessary to store all the raw data and to provide useful patterns with respective to the user needs. Generally, the storage of all raw data will be done in a database maintained by concerned organizations. Data mining

techniques are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To full fill these goals many data mining techniques exists such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

In this process we consider the centralized database are divided into distributed database DB1, DB2,…..DBn and their own key values K_1,K_2,……..K_n, or Select N number of sites each having their own database DB1, DB2,…., DBn . Each site calculates their frequent items set and negative support value.

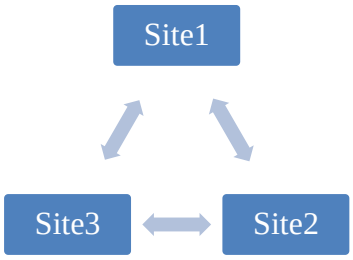

Figure 1: Communication among three sites and DM

Each site are arrange in ring architecture then find the negative partial support , Now the site1 send their negative Partial Support (PS) value to site2 and site2 send their value to site3 and this process continue till site n and after that sites n send their value to site1. Site1 subtract all the Random number value from the Partial Support value and calculate their actual support, now site1 broadcast the actual support value to the entire site present in the distributed environment.

Proposed Algorithm:

Input: Centralized database (Iris Database)

Process: Conversion distributed database, Partial support value, encryption and decryption.

Output: Global Value

Encryption Process

Here number of database server is more than two.

Step1: Take the centralized Database (Iris Database)

Step2: Convert into the horizontally partitioned distributed database (N Number of datasets)

Step3: Calculate the support count of each database.

Step4: Calculate the support and confidence.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Support = $\left(\frac{X \cup Y}{T(total\ number\ of\ transaction)} \right)$

Confidence = $Prob(\frac{X \cup Y}{X})$

Step5: Calculate partial support and partial confidence.

Partial Support (**PS**) = X. Support - DB× Minimum Support

Partial Confidence (**PC**) = X. Confidence - DB× Minimum Confidence

Step6: Add their own private key in all partial support and partial confidence.

Partial Support (**PS**) = X. support - DB×minimum support + Key

Partial Confidence (**PC**) = X. Confidence - DB×Minimum Confidence $+Key$

Step7: Divided the partial support and partial confidence into the three different values.

Step8: converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

Step9: Take the transpose of the matrix (Y^T).

Step10: Convert ASCII code matrix (Y^T) into the binary format.

Step11: Consider our own secret key(X matrix)

Step 12: Covert the X matrix into binary format

Step 13: Perform Exclusive-or between X and Y.

Step14: The resultant matrix is the encrypted format of plain text stored into the associative memory.

Setp15: The resultant matrix is sanded to the protocol initiator Server.

Decryption Process

Step 1: Consider the resultant matrix M

Step 2: Compute transpose of M matrix as M^T matrix

Step 3: Convert matrix M^T into binary format

Step 4: Consider our own private key X

Step 5: Covert matrix X into binary number format

Step 6: Perform exclusive-or operation between M^T and X

Step 7: The resultant matrix is converted to the ASCII code and finally we have the original text.

Step8: After receiving all the original values from the different database, the protocol initiator takes the step for data analysis by calculating Global support and global confidence.

Global Support (**AS**) = $(\sum_{i=1}^{n} \neg \text{PS i} - \sum_{i=1}^{n} \text{Ri})$

Global Confidence (**AC**) = $(\sum_{i=1}^{n} \neg \text{PC i} - \sum_{i=1}^{n} \text{Ri})$

Step9: After that the protocol initiator broadcast the results to all the database server admin presents in the distributed environments.

Step10:- End of the process.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

In this figure 2, we show the analyzed dataset after applying the apriori algorithm in weka tool.

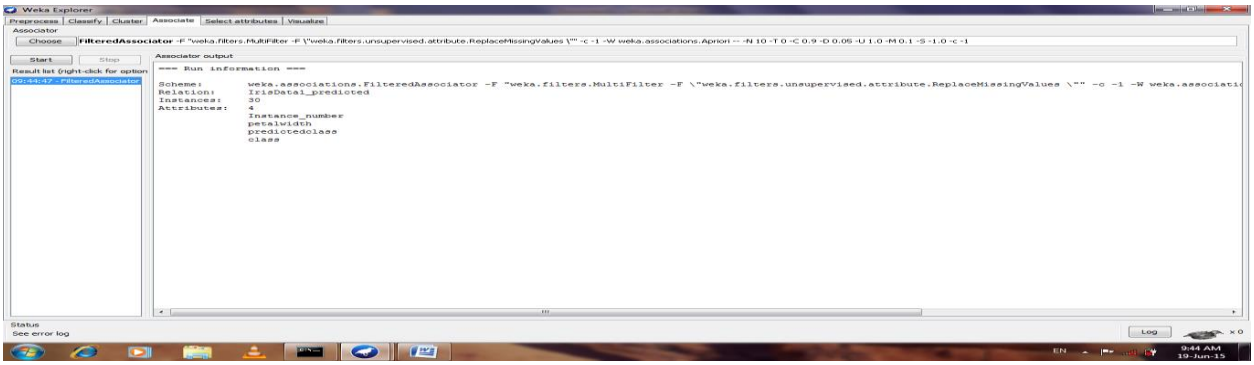

Figure 2: Shows analyzed dataset for the site 1

In this figure 3, we show the analyzed dataset after applying the apriori algorithm in weka tool.

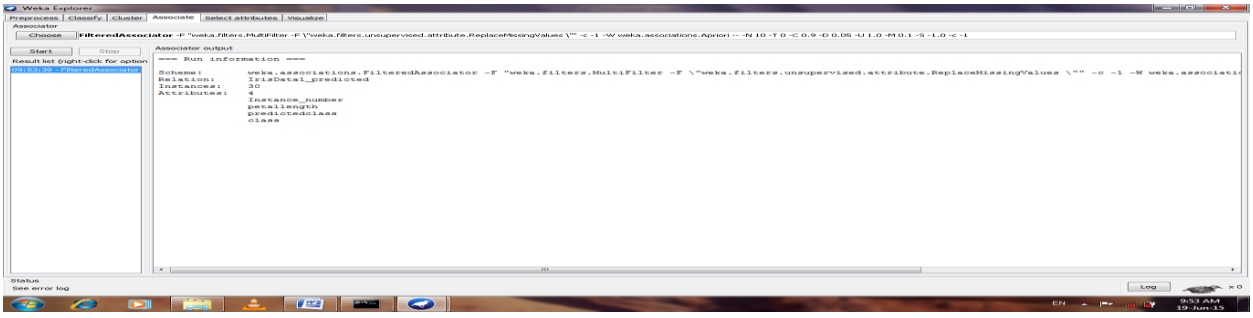

Figure 3: Shows analyzed dataset for the site 2

In this figure 4, we show the analyzed dataset after applying the apriori algorithm in weka tool.

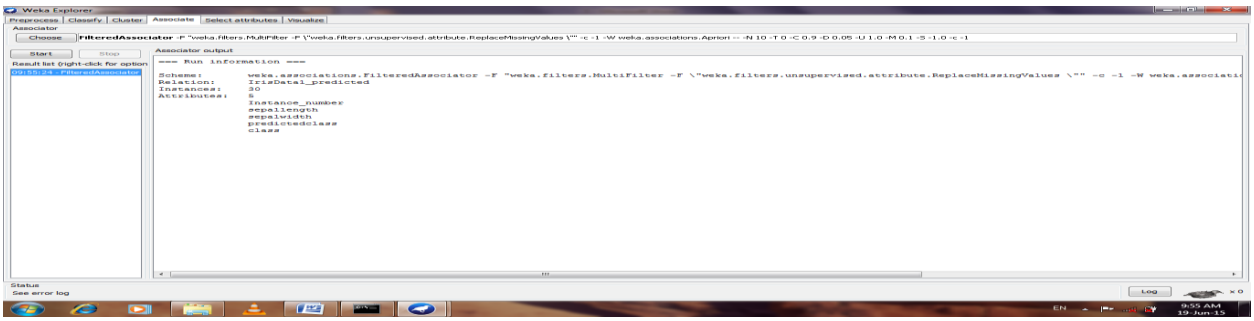

Figure 4: Shows analyzed dataset for the site 3

In this thesis, we take the iris database, first the iris database is centralized then after that, we divided the centralized database into the distributed database, so here we divided it into three different databases and each databases, we represented by different sites, so we have three number of sites. For analyzing the database, we consider the minimum support value is 40% and for providing the highest privacy to centralized database, we consider three random numbers, whose value prospectively 1, 2 and 3. So first each sites calculated their support count by using the Apriori algorithm, then after that each sites calculated their negative partial support by using the formula then after that added the random number and sanded to the next site presented in the distributed environment, than after that the protocol initiator site will calculated their actual support, and broadcast it to all the sites.

Site1:

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

DB=0.2, Key = 1, Support=0.566, Confidence=1

Partial Support (PS) = X. support - DB× Minimum Support + Key

Partial Support (PS) =0.566-(0.2×0.4) + 1=1.486

Partial Confidence (PC) = X. Confidence - DB×Minimum Confidence $+Key$ =17-30*0.6+1=0.0

Divided the partial support and confidence value into the three different values

PS_1=0.486, PS_2=0.50, PS_3=0.50

PC_1=0.0, PC_2=0.0, PC_3=0.0

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

Y= [NULL SOH SOH]

Take the transpose of the matrix (Y^T)

$$Y^T=\begin{bmatrix} NULL \\ SOH \\ SOH \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T=\begin{bmatrix} 000 \\ 001 \\ 001 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_1=\begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}$$

Covert the X matrix into binary format

$$X_1=\begin{bmatrix} 010 \\ 100 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_1=X \text{ Ex-OR } Y^T=\begin{bmatrix} 000 \\ 001 \\ 001 \end{bmatrix} Ex-OR \begin{bmatrix} 010 \\ 100 \\ 001 \end{bmatrix} = \begin{bmatrix} 010 \\ 101 \\ 000 \end{bmatrix}$$

Site2:

DB=1.4, Key= 2, Support = 0.466, Confidence=1

Partial Support (PS) = X. support - (DB× Minimum Support) + Key=0.466-(1.4×0.4) + 2=1.906

Partial Confidence (PC) = X. Confidence - DB×Minimum Confidence $+Key$ =17-30*0.6+1=0.0

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Divided the partial support and confidence value into the three different values

PS_1=0.906, PS_2=0.50, PS_3=0.50

PC_1=0.0, PC_2=0.0, PC_3=0.0

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

Y= [SOH SOH SOH]

Take the transpose of the matrix (Y^T)

$$Y^T=\begin{bmatrix} SOH \\ SOH \\ SOH \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T=\begin{bmatrix} 001 \\ 001 \\ 001 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_2=\begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix}$$

Covert the X matrix into binary format

$$X_2=\begin{bmatrix} 011 \\ 100 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_2=X \text{ Ex-OR } Y^T=\begin{bmatrix} 001 \\ 001 \\ 001 \end{bmatrix} Ex-OR \begin{bmatrix} 011 \\ 100 \\ 001 \end{bmatrix} = \begin{bmatrix} 010 \\ 101 \\ 000 \end{bmatrix}$$

Site3:

DB= 3.4, Key Value= 3, Support= 0.5, Confidence=0.13

Partial Support (PS) =X. support -DB× Minimum Support +Key= 0.5-(3.4×0.4) + 3=2.14

Partial Confidence (PC) = X. Confidence - DB×Minimum Confidence $+Key$ =2-30*0.6+3=0.0

Divided the partial support and confidence value into the three different values

PS_1=0.14, PS_2=1.00, PS_3=1.00

PC_1=0.0, PC_2=0.0, PC_3=0.0

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

Y= [SOH SOH NULL]

Take the transpose of the matrix (Y^T)

$$Y^T=\begin{bmatrix} SOH \\ SOH \\ NULL \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T=\begin{bmatrix} 001 \\ 001 \\ 000 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_3=\begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix}$$

Covert the X matrix into binary format

$$X_3=\begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_3=X \text{ Ex-OR } Y^T=\begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix} \square\square-\square\square \begin{bmatrix} 001 \\ 001 \\ 000 \end{bmatrix} = \begin{bmatrix} 000 \\ 100 \\ 001 \end{bmatrix}$$

After the encryption, now all the sites send their encrypted value with the key value sanded to the protocol initiator site, then the protocol initiator site decrypted that value by using some of the decryption steps that shown in the above algorithm. Perform the Ex-OR operation between the resulting matrix Z_1 Ex-OR X_1, Z_2 Ex-OR X_2 and Z_3 Ex-OR X_3. Then we have the matrix M_1, M_2 and M_3, after that for calculating the resulting matrix M, perform the Ex-OR operation between M_1, M_2 and M_3.

$$M=M_1 \text{ Ex-OR } M_2 \text{ Ex-OR } M_3=\begin{bmatrix} 000 \\ 001 \\ 000 \end{bmatrix}$$

Then take the transpose of the resulting matrix

M^T= [000 001 000] =0.5

After taking the transpose converted into the ASCII value and then we have the value of the resulting matrix M is same as the global support value, If the global support value greater than zero then it means

that the, attribute value that has been taken is globally frequent attribute, it may be locally infrequent attribute. So here in this thesis the calculated value of global support is greater than zero. So it globally accepted.

3. Conclusion

The difficulty of preserving privacy in association rule mining is extracted when the database is distributed horizontally partitioned database among n (n>2) number of sites when no trusted party is considered. In this thesis cryptography algorithm by using the random number is adopted to enhance the privacy further. The proposed replica capably to find global frequent item sets even when no site can be treated as trusted. The trusted party initiates the process and prepares the merged list. All the sites computes the partial supports and total supports for all the item sets in the merged list using the cryptography technique and based on these results finally trusted party finds actual frequent item sets. And after comparing the result of these, we find that data leakage with trusted party is more as compare to without trusted party so privacy is also increased respecting.

References

[1] Agarwal, R., et al "Mining association rules between sets of items in large database". In: Proc. of ACM SIGMOD'93, D.C, ACM Press, Washington, pp.207-216, 1993.

[2]. Agarwal, R., Imielinski, T., Swamy, A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, 1993.

[3]. Srikant, R., Agarwal, R "Mining generalized association rules", In: VLDB'95, pp.479-488, 1994.

[4]Agrawal, R., Srikant, R, "Privacy-Preserving Data Mining", In: proceedings of the 2000 ACM SIGMOD on management of data, pp. 439-450, 2000.

[5] Lindell, Y., Pinkas, B, "Privacy preserving data mining", In: Proceedings of 20th Annual International Cryptology Conference (CRYPTO), 2000.

[6]Kantarcioglu, M., Clifto, C, "Privacy-Preserving distributed mining of association rules on horizontally partitioned data", In IEEE Transactions on Knowledge and Data Engineering Journal, IEEE Press, Vol 16(9), pp.1026-1037, 2004.

[7] Han, J. Kamber, M, "Data mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.

[8]Sheikh, R., Kumar, B., Mishra, D, K, "A Distributed k- Secure sum Protocol for Secure Multi Site Computations".Journal of Computing, Vol 2, pp.239-243, 2010.

[9]Sheikh, R., Kumar, B., Mishra, D, K, "A modified Ck Secure sum protocol for multi partycomputataion".Journal of Computing, Vol 2, pp.62-66, 2010.

[10]Jangde,P., Chandel, G, S., Mishra, D, K.,.: 'Hybrid Technique for Secure Sum Protocol' World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 vol 1, No. 5,198-201, (2011).

[11]Sugumar, Jayakumar, R., Rengarajan, C (2012) "Design a Secure Multi Site Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105.

[12] N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.

[13] N V Muthulakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.

[14]J. Vaidya, "Privacy preserving data mining over vertically partitioneddata," Ph.D. dissertation, Purdue University, 2004.

[15] J. Vaidya, C. Clifton, M. Kantarcioglu, and A. S. Patterson, "Privacypreservingdecision trees over vertically partitioned data," in ACMTransactions on Knowledge Discovery from Data, vol. 2, no. 3, 2008,pp. 14–41.

[16] Y. Shen, H. Shao, and L. Yang, "Privacy preserving c4.5 algorithmover vertically distributed datasets," in International Conference onNetworks Security, Wireless Communications and Trusted Computing,vol. 2. Wuhan, Hubei: IEEE computer society, April 2009, pp. 446–448.

[17] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mentalgame or a completeness theorem for protocols (extended majority36abstract)," in STOC '87 Proceedings of the nineteenth annual ACMsymposium on Theory of computing, New York, 1987, pp. 218–229.