Journal of Engineering and Applied Sciences 13 (17): 7187-7191, 2018
ISSN: 1816-949X

Fast and Secure Association Rule Mining on Distributed Databases Using FDM and RSA Algorithms

[1]J. Sumithra Devi and [2]M. Ramakrishnan
[1]Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]School of Information Technology, Madurai Kamaraj University, Madurai, Tamil Nadu, India

Abstract: In association rule mining, leakage of sensitive data can cause potential threats to privacy and data protection. In distributed database architecture, performing association rule mining following traditional privacy preserving techniques are not feasible. We present a novel privacy preserving association rule mining algorithm that uses cryptosystem technique to maintain privacy. We use FDM technique to find frequent itemsets. The support count is encrypted using RSA algorithm and forwarded to other sites. We use one data initiator, one data combiner and other parties as client in ARM process. Experimental results show that this method is flexible and ensures privacy during global support count calculation process.

Key words: Association rule mining, privacy preservation, RSA, data mining security, distributed data mining, mining

INTRODUCTION

Data mining is one of the most challenging research area that digs and analyzes enormous amount of data and extracts useful information from them. It is a computer assisted process that predicts behaviour of the current system and future trends that make business people to take aggressive, knowledge driven business decisions (Castro *et al*., 2007). Hence, data mining is not a forbearing collection of data it is a powerful strategy, so, important in data analytics, exploring trends and new finding business patterns. Data mining process is composed of several important and overlapping techniques such as association, classification, clustering, prediction, sequential patterns, decision making, etc., (Romero and Ventura, 2010).

Among the above techniques, association is a best known data mining technique which discovers relationship between items in the same transaction (Tajbakhsh *et al*., 2009). Association rule mining, abbreviated as ARM, tries to find frequent co-occurring associations among collection of data items. The idea is to find the association of items that more often appears together in almost random sampling of all possibilities (Hastie *et al*., 2009). This process is sometimes referred to as market basket analysis.

Because of globalization, there is an enormous expansion of business activities. Many organizations are operating their business in several locations. Each location is having their own private data and this forms the distributed database architecture (Coulouris *et al*., 2005). In order to perform association rule mining process on these distributed databases it is mandatory that all the parties need to share their confidential data to the ARM system to perform. This leads to the issue of maintaining privacy during ARM process.

It is customary to provide privacy to any mining system, so that, sensitive data should not be leaked (Dwork, 2008). Existing methods that provides privacy during ARM on distributed private databases tries to compute the answer without revealing any information about the user. However, this approach fails to provide cumulative answer for the ARM. Third party authentication is another technique which gathers data from several main parties and third party performs ARM (LeMay *et al*., 2007). Again, all the main parties has to completely trust the third party and it is also, not that much correct to believe on one party under the current security breaching environments (Zissis and Lekkas, 2012).

Another method to provide privacy is secure multi-party computation where two parties calculate their output using a function and output for the whole system is calculated in this method (Malkhi *et al*., 2004). This method, also, contains certain drawbacks as huge computation and time is involved. Hence, in this study, we present a novel privacy preserving association rule mining using cryptographic techniques for distributed databases.

Literature review: Kantarcioglue *et al*. (2009) presented an efficient and approximate protocol for privacy preserving during association rule mining process. Secure scalar product (dot product) with minimal computation

Corresponding Author: J. Sumithra Deviand, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

cost on large vectors is provided. The ARM is performed on vertically partitioned data. Different ways are presented to efficiently compute the dot product approximation.

Pathak *et al.* (2012) proposed privacy preserving association rule mining using impact factor concept. It is a heuristic based association rule mining algorithm, uses impact factor of the transactions. The impact factor of a transaction is the number of itemsets that are present in sensitive association rules. This method modifies fewer transactions thereby maintaining data quality.

Yi *et al.* (2015) proposed privacy preserving association rule mining in cloud computing. In order to overcome the high computation cost of k-anonymity, k-support and k-privacy techniques, user encrypts its data and stores it in the cloud and mining of association rules are done using outsourced semi-honest servers. These semi honest servers perform association rule mining on encrypted data and outputs association rules to the user. This algorithm provides distributed cryptosystem to achieve privacy, transaction privacy and database privacy. Compromisation of all servers is tackled by selecting servers from different cloud servers.

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

Zhu and Li (2015) proposed a privacy preserving association rule mining scheme based on Hybrid Partial Hiding (HPH) strategy. The original dataset is altered and changed to different random parameters. The frequent itemsets are generated with altered data using hybrid partial hiding strategy. Usage of HPH algorithm improves the privacy preservation.

Chandrakar *et al.* (2010) proposed privacy preserving association rule mining method using hybrid algorithm. Sensitive information used for association rule mining is altered by combining the increase support of Left Hand Side (ISL) and decrease support of right hand side (DSR) methods. This increase/decrease of support count maintain privacy during association rule mining process. The results show that this method is secure, prunes more number of sensitive rules with minimum number of database scans.

From the above concise literature review it is clear that already enough research is going on in maintaining privacy during association rule mining process.

Background: Association rule mining process generally finds repeated co-occurring association of a collection of items in a transaction. This collection is termed as frequent itemset which produces significant association rules. Association rule is represented with an implication of the form $X \Rightarrow Y$ where X and Y are often occurring itemsets in the transaction database (Wu *et al.*, 2008). The association rule consists of 2 parts; The left hand side and the right hand side or antecedent and consequent respectively (Verykios *et al.*, 2004). The association rule has the support 'S' such that S% of transactions contains $X \cup Y$. It also, holds confidence 'C' such that C% of transactions that contain X, also, contains Y.

Distributed association rule mining: It is possible that an enterprise may have different franchise selling the enterprise product in different locations. Each franchise is having their own dataset, maintaining their transactions. If the enterprise wants to perform association rule mining, it has to collect confidential data from all of its franchise. Here, comes the problem of maintaining privacy as the franchises want to keep their transaction data undisclosed with other franchises. This problem is said to be privacy preserving association rule mining on distributed data.

Let D be the database containing N transactions. Assume that 'n' sites are available such that $S_1, S_2, S_3, \ldots, S_n$ in a distributed fashion and all the database is partitioned over 'n' sites such as $D_1, D_2, \ldots, D_n$. Let X_{sup} and X_{supi} be the support counts of an itemset X in D and D_i. X_{sup} is the global support count and X_{supi} is the local support count of X at site S_i. The global support and confidence is calculated as:

$$\text{Global support} = S_{G(X,Y)} = \frac{\sum_{i=1}^{n} \text{Support}_{\text{count}_{X,Y}}(i)}{N}$$

$$\text{Global confidence} = \text{C-G.} = \frac{\text{Support}_{G(X,Y)}}{\text{Support}(X,Y)}$$

To have a secure distributed model it is customary to check whether each potential itemset satisfies threshold values. For this, summation and comparison protocols are used that tests a particular itemset is globally supported or not. Here, each site computes local support for an itemset. Then site uses secure summation technique which adds random value 'R' to the excess support count and pass it to the next site. The next site adds its excess support count value and passes it to the next. When the first site gets back the support count it subtracts its random value R to get the actual global support value. The last site performs a comparison with first site to check if support $\geq$R.

The distributed algorithm requests all the sites to send the rules whose support count is atleast 'K' where K is the user specified support. For each rule supplied by the sites, requests all the sites to send their support count

for that rule and total count of all transactions at the site. Global support for each rule is computed and rules with minimum support count can be grouped.

MATERIALS AND METHODS

Fast Distributed Mining (FDM) uses effective pruning strategies and generates fewer candidate itemsets in association rule mining process. For each site, FDM locates itemsets with local support count and prunes infrequent itemsets. FDM sends the local counts to polling site instead of broadcasting that minimizes the communication overhead. FDM generates fewer candidate itemsets even on heterogeneous datasets. Hence, we use FDM in our proposed method to perform association rule mining.

To perform FDM on datasets, sites are classified as data combiner, data initiator and clients. In a simulation, one party is data combiner, another one party is data initiator and remaining parties are clients to the data combiner. One round is used for computing frequent itemsets and global support. The mining algorithm takes less time as it uses minimum scan over data.

Each site computes the candidate itemset using FDM and itemsets that have support count above the threshold value are grouped together and termed as LLi(K) which are locally large itemsets. This group is encrypted using any cryptosystem along with its support count and termed as LLei (K). This is communicated to data combiner.

The data combiner combines all the received frequent itemsets and support counts with its own frequent items and support count which is also in encrypted form. This is then communicated to data initiator to compute the global association rules. The data initiator receives encrypted frequent itemset with support, decrypts it and merges with local data mining results to obtain global mining results. The result is passed to all the sites participated in association rule mining process.

RSA algorithm; Mining process:

```
Procedure arm (N, X, Y, K)
{
   // N→ number of sites participating in ARM & N>3
   // X, Y → list of items
  // K → minimum support threshold

      Perform frequent itemset mining
      Encrypt all itemsets and its support count

for each site
{
      generate LLi(K) using FDM
      LLei(K) = ϕ
      for each itemset∈LLi(K)
      {
        calculate local support
        LLei(K) = LLei(K)∪{Ei(X), Ei(Y)}
      }
      end for
}
end for

for each rule
      compute LLi(K) using FDM
      for each X ∈ LLi(K)
      {
        calculate local support of X and Y
        LLei(K) = LLei(K) {Ei(X), Ei(Y)}
      }
      end for
 add all the values of LLe(K)
 decrypt LLe(K) and compute LLi(K) to compute global support

 for each itemset ∈ LLe(K)
 {
   LL(K) = D(LLe(K))
   compute local support of X
 }
 end for

end arm
```

The algorithm works in 3 steps regardless of any number of clients in calculating global candidate itemsets. In RSA algorithm, private and public keys are used. Let K_1 represents private key and K_2 represents public key. Initially there may not be any communication between clients and results are encrypted. Since, data combiner is unaware of private key, there is no privacy breach.

Since, each site's results are mixed, data initiator cannot connect any data and site, hence, there is also, no privacy breach. Even when data initiator computes final L(K) and publish association rules, no site can conclude other site's data. So, this method is secure and privacy preserving.

RESULTS AND DISCUSSION

In order to evaluate the performance of the proposed method, we calculate computation cost, communication cost and accuracy. Let 'N' be the number of sites participating in association rule mining. LLi(K) be the number of frequent itemset and let L(K) be the number of frequent itemsets whose support count is higher than threshold value. For an itemset, let 't' represents number of bits after encryption. The communication cost of the proposed method is given by:

$$0\left(t^2 * | L(K) * N *\right)$$

Since, the evaluation is on distributed association rule mining, we use different sites having data and remote invocation method to connect sites. The data initiator

Table 1: Computation time for various data sizes

Data size	Multi party method	Our method
2 KB	0.1347	0.0984
25 KB	1.5731	0.6748
50 KB	2.7258	0.9244
100 KB	5.6683	1.3847
1 MB	11.5834	4.4531
2 MB	110.2863	58.6797

Table 2: Comparison on number of rules generated

Data size	Multi party method	Our method
2 KB	153	153
25 KB	146	146
50 KB	235	235
100 KB	668	668
1 MB	986	986
2 MB	1,358	1,358

defines the support and confidence values, generates private and public keys and encrypts the data using its public key. The data combiner combines results of different sites and accumulates the result. The experiment was conducted on 0/1 matrix data. The size of the data varies from 2 KB-2 MB. Experiment was conducted on a computer with 2.8 GHz processor and with 512 MB RAM. Table 1 provides computation time of our method compared to traditional method.

Regardless of any method, the number of genuine association rules must be same. To test the accuracy of the proposed method, we use the same datasets and the results in terms of number of rules generated are tabulated in Table 2. We use support count as 40% and confidence as 60%.

From the above experimental result, it is clear that the proposed method shows good performance in terms of computation time, communication time and accuracy. Since, any number of sites can be added without any change in implementation, this method is highly flexible. Moreover, addition of sites cannot increase the computation time because each site is individually calculates its support count. This algorithm generates frequent itemsets and local support count simultaneously.

CONCLUSION

In this study, privacy preserving association rule mining in distributed environment is presented. We use FDM algorithm to mine frequent itemsets and RSA algorithm to encrypt the data. We use data initiator that starts the SRM process, data combiner that combines local support count of different sites. All the sites which send local support count are clients. Since, the data are encrypted with public key of data initiator it alone can view the data. There is no security breach in the algorithm. This method is flexible, accurate and good in performance. One of the significant features of this method is that any number of sites can be added in ARM process without adjusting the algorithm.

REFERENCES

Castro, F., A. Vellido, A. Nebot and F. Mugica, 2007. Applying Data Mining Techniques to E-Learning Problems. In: Evolution of Teaching and Learning Paradigms in Intelligent Environment, Jain, L.C., R.A. Tedman and D.K. Tedman (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-71973-1, pp: 183-221.

Chandrakar, I., Y.U. Rani, M. Manasa and K. Renuka, 2010. Hybrid algorithm for privacy preserving association rule mining. J. Comput. Sci., 6: 1494-1498.

Coulouris, G.F., J. Dollimore and T. Kindberg, 2005. Distributed Systems: Concepts and Design. Addison-Wesley, Boston, Massachusetts, USA., ISBN:9780321263544, Pages: 927.

Dwork, C., 2008. Differential Privacy: A Survey of Results. In: Theory and Applications of Models of Computation, Agrawal, M., D. Du, Z. Duan and A. Li (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-79227-7, pp: 1-19.

Hastie, T., R. Tibshirani and J. Friedman, 2009. Unsupervised Learning. In: The Elements of Statistical Learning, Hastie, T., R. Tibshirani and J. Friedman (Eds.). Springer, New York, USA., ISBN:978-0-387-84857-0, pp: 485-585.

Kantarcioglu, M., R. Nix and J. Vaidya, 2009. An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining. In: Advances in Knowledge Discovery and Data Mining, Theeramunkong, T., B. Kijsirikul, N. Cercone and T.B. Ho (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-01306-5, pp: 515-524.

LeMay, M., G. Gross, C.A. Gunter and S. Garg, 2007. Unified architecture for large-scale attested metering. Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), January 3-6, 2007, IEEE, Waikoloa Village, Hawaii, pp: 115-115.

Malkhi, D., N. Nisan, B. Pinkas and Y. Sella, 2004. Fairplay-secure two-party computation system. Proceedings of the 13th Symposium on USENIX Security Vol. 4, August 9-13, 2004, USENIX, San Diego, California, pp: 1-17.

Modi, C.N. and A.R. Patil, 2016. Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Without Involving Trusted Third Party (TTP). In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, Nagar, A., D. Mohapatra and N. Chaki (Eds.). Springer, India, ISBN:978-81-322-2528-7, pp: 549-555.

Pathak, K., N.S. Chaudhari and A. Tiwari, 2012. Privacy preserving association rule mining by introducing concept of impact factor. Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA'12), July 18-20, 2012, IEEE, Singapore, ISBN:978-1-4577-2118-2, pp: 1458-1461.

Romero, C. and S. Ventura, 2010. Educational data mining: A review of the state of the art. IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev., 40: 601-618.

Tajbakhsh, A., M. Rahmati and A. Mirzaei, 2009. Intrusion detection using fuzzy association rules. Appl. Soft Comput., 9: 462-469.

Verykios, V.S., A.K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, 2004. Association rule hiding. IEEE. Trans. Knowl. Data Eng., 16: 434-447.

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh and Q. Yang *et al.*, 2008. Top 10 algorithms in data mining. Knowledge Inform. Syst., 14: 1-37.

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, ACM, Singapore, ISBN:978-1-4503-3245-3, pp: 439-450.

Zhu, J. and Z. Li, 2015. Privacy Preserving Association Rule Mining Algorithm Based on Hybrid Partial Hiding Strategy. In: LISS, Zhang, R., Z. Zhang, K. Liu and J. Zhang (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-40660-7, pp: 1065-1070.

Zissis, D. and D. Lekkas, 2012. Addressing cloud computing security issues. Future Generation Comput. Syst., 28: 583-592.