International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Improving Healthcare using Privacy Preserving Association Rule Mining in Distributed Healthcare Data

Nikunj H. Domadiya, Arpesh Kumar, Udai Pratap Rao

Abstract*: The trend of data mining in healthcare has increased due to the digitalization of hospitals with electronic health records (EHR) system. The data stored at EHR systems are valuable assets for medical research. Association between disease and patient's symptoms facilitate the doctors in taking healthcare decisions. The accuracy of decision can be enhanced by association rule mining on distributed healthcare data. The prerequisite is to share healthcare data by all collaborative EHR systems. Disclosing patient's healthcare data for collaborative data mining may cause privacy issues. Privacy preserving distributed data mining solves this problem by achieving privacy and collaborative data mining results. Existing cryptography solutions provide privacy with higher computation and communication cost. In this paper, we propose an efficient approach for finding association rules in distributed horizontally partition healthcare data with comparatively efficient communication and computation cost. The theoretical and practical evaluation shows that our approach is efficient, scalable and outperforms existing approach. At last, we have shown the real application of proposed approach for breast cancer prediction based on symptoms of patients.

***Index Terms*: Distributed Healthcare Data, Association Rule Mining, Privacy Issues, Privacy Preserving Data Mining, Breast Cancer Prediction.**

I. INTRODUCTION

Currently, major healthcare organizations store patient's data in the digital form using electronic healthcare record (EHR) system [1]. Knowledge discovery techniques extract useful patterns from the collected data to support healthcare decisions. Association between some sensitive disease and its symptoms, treatment, and medicine helps in predicting the disease in early stage [2][3]. Early prediction of these diseases reduces the mortality rate and save more human lives. Association rule mining technique is very useful in extracting patterns between disease and patient's symptoms with specific precision. The precision of healthcare decision can be improved by aggregating the healthcare data stored in EHR systems at different location by different healthcare providers. Privacy becomes an important concern while publishing the patient's data during the aggregation of healthcare data [1]. Hence, medical research majorly focused on distributed data mining and privacy preservation. Privacy preserving distributed data mining (PPDDM) becomes a prominent research topic in the area of privacy and healthcare data mining. There are two major types of partition in distributed healthcare data, viz, horizontally and vertically partitioned data. In horizontally partition data, all the participants have same schema and different number of transactions while in vertically partition database, all the participants have different schema and equal number of transactions.

There are two models of PPDDM viz. the World Wide Web model and the corporate model [2]. In corporate model, we assume that each party has data stored in a specific format, while in the world wide web model, we use the individual provided data. A classical data mining will not be used, when parties want to share the data and perform the collaborative data mining because of some privacy law and policy. Therefore, design of an efficient algorithm is required in which participants may be allowed to collaboratively perform the data mining while maintaining individual privacy. There are mainly two approaches of privacy preserving data mining (PPDM) viz, *randomization approach* and *cryptography approach*. Cryptography approaches [1][3][4] [5][6][9] provide higher privacy but suffer from high computation and communication complexity. Hence, the existing techniques are not scalable in terms of database size and number of participants. Therefore, it is important to design a new algorithm which is scalable and provide high level of individual privacy with lower computation and communication cost. In major healthcare providers, EHR systems follow the same schema while storing the patient's data. Hence, we focused on horizontally partitioned data. In this paper, we propose a non-cryptography approach for privacy preserving association rule mining from horizontally partition database. In the existing literature, some techniques [7][8][21] proposed secure mining of association rules in horizontally partition data which are based on the FDM algorithm [9] and set union problem related to the threshold function. These techniques used the threshold function and set inclusion computation, but their communication and computational cost is very high, and it seems to be in terms of cube [6] and square [8] because of their message broadcasting nature. Therefore, to overcome this problem, we propose an efficient approach for mining global association rules in horizontally partition databases using the randomized communication path technique.

Manuscript published on 30 April 2019.
* Correspondence Author (s)
Nikunj H. Domadiya*, S.V.N.I.T, Surat, India.
Arpesh Kumar, S.V.N.I.T, Surat, India.
Udai Pratap Rao, S.V.N.I.T, Surat, India.

A. Problem Description

Now a day, preserving the privacy of individuals is the challenging task for data mining research community. Most of the data are distributed at a different location and aggregating them at single location violates the privacy policies. Therefore, privacy preserving distributed data mining is essential in distributed data mining. The secure multiparty computation (SMC) [10] has been introduced to deal with privacy preserving distributed data mining [11]. However, the existing cryptography based approaches of SMC incur higher computational as well as communication overhead [12]. Moreover, these algorithms have strict assumptions of the collaborative participants as they will not collide with each other. Hence, it is essential to design an approach which is better compared to the existing approaches in terms of communication and computational cost as well as security of which does not rely on the assumption of trusted or non-colluding parties.

We have considered the EHR system as a collaborative participant who provides the data and interested in global association rule mining for mutual benefits. All collaborative participants follow the algorithm and access the global association rule over aggregated data while preserving the privacy of an individual. Propose approach is based on a randomized communication path among the participants.

II. THEORETICAL BACKGROUNDS

Two or more participants collaborate with each other for some mutual benefits by sharing their information or data. Each participant is interested in aggregate mining result without revealing own information. This problem can be solved by secure multiparty computation (SMC). Gold wasser [13] outlined the SMC problem in distributed database. There are several solutions of privacy preserving association rule mining. However, some algorithms depend on trusted third party, where all participants send their data to trusted third party. Trusted third party finds the global mining results and share global results to all participants. In real environment, finding a trusted third party is difficult. In semi-honest model, without trusted third party, all participants find the global results and maintain the privacy of individual participants. We have considered the semi-honest model for our approach.

A. Association Rules Mining in Horizontally Partitioned Databases

The standard definition of association rule in distributed environment (on horizontal partitioned data) is given as follows [14]: Let $I = \{i_1, i_2, i_3 \ .., i_m\}$ be a set of items. Database D is a set of transactions, which is distributed among n sites named $\{P_1, P_2, ..,P_N\}$ in such a way that database D $(1 \le i \le n)$ maintained by participant P consists of same set of items but different number of transactions. The itemset $X \subseteq I$ has local support count $l_i(X)$ at site P_i. The global support count of X is given as $g(X) = \sum_{i=1}^{n} l_i(X)$. An itemset X is globally supported if $g(X) \ge$ *Minimum support threshold.* The efficient algorithms for mining distributed association rule are given in [14]. Association rule XY→Z can be computed using following equations:

$$Support(XY \rightarrow Z) = \frac{\sum_{i=1}^{sites} support\ count\ of\ \ XYZ\ at\ site_i}{\sum_{i=1}^{sites} size\ of\ \ Database\ at\ site_i} \quad (1)$$

$$Support(XY) = \frac{\sum_{i=1}^{sites} support\ count\ of\ \ XY\ at\ site_i}{\sum_{i=1}^{sites} size\ of\ \ Database\ at\ site_i} \quad (2)$$

$$Confidence(XY \rightarrow Z) = \frac{support(XY \rightarrow Z)}{support(XY)} \quad (3)$$

B. Related Work

Tamir Tassa [8] proposed a secure protocol for mining of association rules in horizontally partitioned data. He had modified the work of Murat et. al [6] and considered the threshold set union problem, which is related to the threshold function. It reduces the communication and computational cost significantly as compared to the existing algorithm [6].

Jayanti et al. [14] used randomized response technique for privacy preservation in horizontally partitioned databases. This algorithm works with semi-honest model.

Clifton et al. [2] presented the individual's privacy protection issues in the database. The authors discussed the research issues in the area of privacy preserving data mining techniques [15] [16]. In [17], the authors proposed a protocol for calculating privacy preserving data mining algorithms which help to find the various features according to different criteria of privacy preserving algorithms.

Evfimievski et al. [18] discussed a new protocol for privacy preserving association rule mining. A novel algorithm is discussed in [19], which find privacy preserving association rule mining in centralized database as well as balances knowledge discovery and privacy preserving in association rule mining. Gkoulalas Divanis et al. [20] presented several issues associated with privacy preserving association rule mining.

III. PROPOSED APPROACH

Some of the notations used in algorithms are shown in table I. Distributed association rule mining in horizontally partitioned data is shown in *Algorithm I*. Step 5 of this algorithm requires the sharing of private information by each participant to find the global count of itemset. Proposed approach, shown in *Algorithm II* can be applied to preserve privacy during computation of global count of itemset.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table I: Notation Used in the Proposed Algorithm

Notation	*Explanation*
P	Set of all participants=$\{P_1,P_2,P_3,...,P_n\}$
F	Set of all frequent itemset
F_k	Set of all frequent (k-itemset)
C_k	Set of all candidate (k-itemset)
C_k^i	Frequency vector of all candidate k-itemset with participant P_i

Algorithm 1: Distributed Association Rule Mining

1. $F_1 = find\ frequent\_1\text{-}itemset(D)$
2. ***for*** $(k = 2; F_{k-1} \neq \emptyset; k{+}{+})$
   {
3. $C_k = apriori\_gen(F_{k-1})$;
4. ***for each*** $candidate\ c \in C_k$
5. $collaboratively\ find\ the\ c._{count}\ (use\ the\ algorithm\ 2)$;
6. $F_k = \{c \in C_k \mid c.count \geq min\_sup\}$
   }
7. $return\ F = \cup_k F_k$;

Proposed algorithm can be incorporated with basic distributed frequent itemset mining to preserve privacy of all participants with frequent itemset mining. Proposed approach first builds the random communication path among all participants. Participant P1 becomes an initiator and decides the successor participant of all the participants to form a random path among all participants. Each participant sends the data to only its successor. Due to this random communication path, proposed approach is collision resistant as communication path is changed after each iteration.

Algorithm 2: Proposed Algorithm for Privacy preserving Frequent Itemset mining

1: *All participant P_i computes the C_k^i $(1 \leq i \leq n)$*

2: *Participant P_1 build the random circular communication path among the participants and send successor participant information to all participants.*

3: *P_1 generates and send a random vector to its successor.*

4: ***for*** *i=2 to n*

5: *Each participants P_i add its own vector C_k^i with received vector and send it to successor participant.*

6: ***end for***

7: *Initiator P_1 add its own vector with received vector and subtract the initial random vector from it.*

8: *Initiator computes the final vectore and share it with all the participants*

Privacy preserving computation of candidate itemset count in the step 5 of Algorithm 1 can be achieved using proposed algorithm. Participant P_1 generates a random vector and sends it to its successor participants. Participants P_i receives the vector and add its own vector C_k^i and send it to successor participants. Finally, participant P_1 receives the vector. It subtracts the random vector and adds its own vector C_k^1. P_1 gets the global count of all the candidate k-itemset and publishes it to all the participants. Itemsets whose count is greater than minimum support threshold are included in the list of frequent k-itemsets. Proposed approach preserve the privacy and finds the accurate results. Due to the random communication path among the participants, colluding participants could not infer any private information related to other individual participants. It uses the simple mathematical operations instead of complex cryptographic operation. As a result, proposed approach has less computation and communication cost compared to existing cryptography based approach [6] [8].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Existing algorithms, UNIFI-KC [6] and UNIFI [8] fail to preserve privacy in case of collusion between any two or more participants. Proposed algorithm maintains individual privacy in the presence of collusion among the participants. The use of random shuffle makes difficult for any participants to find the collude participant in each round.

We use synthetic dataset which has been used by Tamir Tassa [8]. The detail of dataset is shown in table II.

Table II: Details of Synthetic Dataset

Number of transactions in whole database	1,00,000
Number of Items	1000

We have analyzed the proposed approach with 5 participants by randomly dividing the original dataset into 5 partitions. We replicated these 1,00,000 transactions to 5,00,000 transactions for analyzing the proposed approach with higher dataset size. We partition the transactions according to the number of participants.

A. Performance Results

Table III shows the comparison between UNIFI [8] and proposed the algorithm in terms of computation time v/s varies number of transaction with fixed number of participants.

Table III: Comparison between UNIFI [8] and Proposed Approach for Total Computation time (in second) v/s Number of Transactions (Number of Participants =5)

N (Number of Transactions)	Total Computation Time	
	UNIFI	Proposed Approach
100000	30.37	26.99
200000	59.24	27.45
300000	87.55	29.48
400000	98.84	30.58
500000	110.12	34.16

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table IV: Comparison between UNIFI [8] and Proposed Approach for Total Computation time (in second) v/s Number of Participants (Number of Transactions = 5, 00,000)

M (Number of Parties)	Total Computation Time(in second)	
	UNIFI	Proposed Approach
5	100.11	20.28
10	111.54	32.14
15	117.34	32.69
20	120.45	34.53
25	123.76	35.96

Table IV shows the comparison between UNIFI [8] and proposed approach in terms of total computation time v/s varies number of participants.

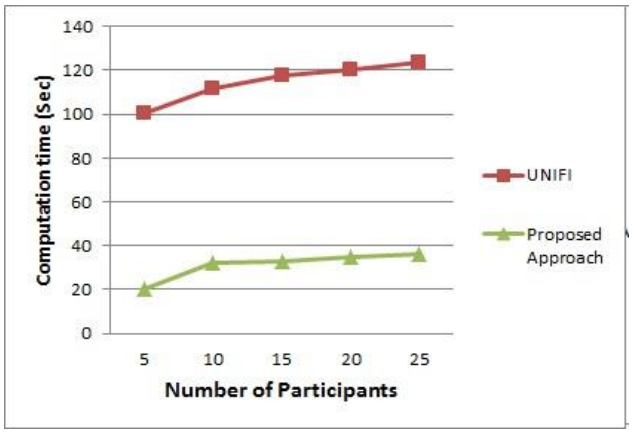

Fig. 1: Comparison between Proposed Approach and UNIFI[8] for Computation time Vs Number of Participant

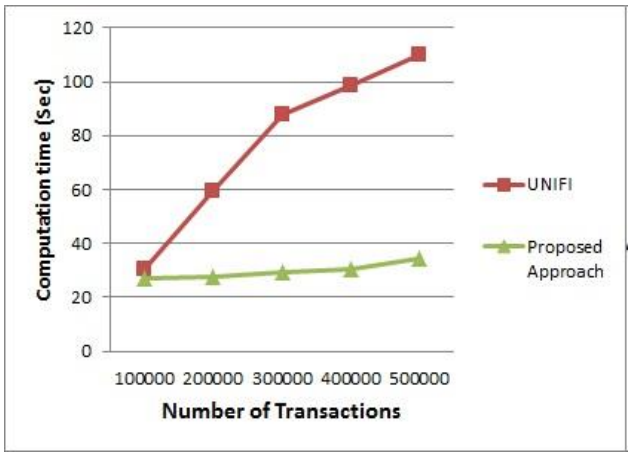

Fig. 2: Comparison between Proposed Approach and UNIFI[8] for Computation Time Vs Number of Transactions

As shown in fig. 1 and fig. 2, proposed algorithm works with lower computational cost compared to existing approach UNIFI [8] with the increasing number of transactions and number of participants. Thus, it is scalable in the real-life scenario.

V. APPLICATION OF PROPOSED APPROACH: IMPROVING PRECISION OF BREAST CANCER PREDICTION

Breast cancer is one of the deadliest diseases in women. We used Wisconsin breast cancer dataset (publicly available on UCI repository [22]) to analyze the proposed approach in a real application in healthcare. We analyze the symptoms and their correlation with breast cancer. We have conceded 4 collaborative healthcare with EHR systems and compared the experimental results of the proposed approach with the results of data mining at any single EHR system by each participant. We have considered 10 out of 32 significant attributes in our analysis. Breast cancer state is defined with two label *malignant* and *benign*.

Association rules with RHS value as *malignant* and *benign* are selected for prediction analysis. The precision (*confidence (%)*) of association rule is improved using the proposed approach compared to any single participant results. Comparison of the proposed approach with single participant results is shown in fig 3. As shown in fig 3, precision of breast cancer prediction at each participant is lower compared to proposed approach. Hence, all participants (healthcare providers with EHR system) get the benefits using the proposed approach as proposed approach allows accessing the global association rules with higher precision while preserving the privacy. It also facilitates extracting the patterns of symptoms and some specific disease to the medical researchers and other physicians. Proposed approach can be applied to analyze the correlation between other critical disease and symptoms.

Association Rules	Accuracy / Confidene (%)				
	EHR1	EHR2	EHR3	EHR4	Proposed Approach
{Bare nuclei=7 ***and*** Single epithelial cell size=1 ***and*** Normal nucleoli=2} --> {Class = benign}	92%	93%	90%	94%	97.50%
{Uniformity of cell shape= 10 ***and*** Uniformity of cell size=8 ***and*** Marginal adhesion=10} --> {Class=malignant}	95%	91%	89%	93%	98%
{Uniformity of cell size=8 ***and*** Clump thickness=7 ***and*** Bland chromatin=1 ***and*** Marginal adhesion=7} --> {Class=malignant}	90%	91%	92%	89%	97%
{Single epithelial cell size=2 ***and*** Uniformity of cell shape=3 ***and*** Bare nuclei=1} --> {Class = benign}	92%	89%	93%	95%	98%
{Bland Chromatin=5 ***and*** Mitoses=3} --> {Class=malignant}	94%	92%	93%	91%	98.50%
{Uniformity of cell shape=2 ***and*** Clump thickness=5} --> {Class=benign}	89%	90%	94%	92%	97.50%

Fig. 3: Breast Cancer Prediction accuracy (confidence (%)) at each EHR system and using proposed approach.

VI. CONCLUSION

Healthcare services can be improved using association rule mining in distributed healthcare data. Privacy becomes an important aspect for distributed healthcare data mining. We proposed an efficient approach for privacy preserving association rules mining in horizontally partition distributed healthcare data. Proposed approach gives better privacy with lower communication and computational cost. Further, experimental analysis shows that our proposed approach is also scalable compared to the existing approach with large datasets and higher number of collaborative participants.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

REFERENCES

1. P. C. Tang and C. J. M Donald, "Electronic health record systems," Biomedical informatis, vol. 10, no. 4, pp. 447-475, 2006.
2. C. Clifton and D. Marks, "Security and privacy implications of data mining," in ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 15–19, 1996.
3. I. S. Alwatban and A. Z. Emam, "Comprehensive survey on privacy preserving association rule mining: Models, approaches, techniques and algorithms," International Journal on Artificial Intelligence Tools, vol. 23, no. 05, pp.145–162, 2014.
4. S. Rana and P. Thilagam, "Hierarchical homomorphic encryption based privacy preserving distributed association rule mining," International Conference on Information Technology (ICIT), pp. 379–385 , 2014.
5. Domadiya, Nikunj, and Udai Pratap Rao. "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data." Procedia computer science , Vol. 148, pp. 303-312, 2019.
6. M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge & Data Engineering, no. 9, pp. 1026–1037, 2004.
7. Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in CryptologyCRYPTO 2000. Springer, 2000, pp. 36–54.
8. T. Tassa, "Secure mining of association rules in horizontally distributed databases," Knowledge and Data Engineering, IEEE Transactions on, vol.26, no. 4, pp. 970–983, 2014.
9. Domadiya, Nikunj, and Udai Pratap Rao. "Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases." Sādhanā Vol. 43, No. 8, pp. 127-132, 2018.
10. A. C.-C. Yao, "Protocols for secure computations," in FOCS, vol. 82, pp. 160–164, 1982.
11. A. Ben-David, N. Nisan, and B. Pinkas, "Fairplaymp: a system for secure multi-party computation," in Proceedings of the 15th ACM conference on Computer and communications security. ACM, pp. 257–266, 2008.
12. D. Beaver, S. Micali, and P. Rogaway, "The round complexity of secure protocols," in Proceedings of the Twenty-second Annual ACM Symposium on Theory of Computing, ser. STOC '90. New York, NY, USA: ACM, pp. 503–513, 1990.
13. S. Goldwasser, "Multi party computations: past and present," in Proceedings of the sixteenth annual ACM symposium on Principles of distributed computing. ACM, pp. 1–6, 1997.
14. J. Dansana, R. Kumar, and D. Dey, "Privacy preservation in horizontally partitioned databases using randomized response technique," in Information & Communication Technologies (ICT), 2013 IEEE Conference on. IEEE, pp. 835–840, 2013.
15. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM Sigmod Record, vol. 33, no. 1, pp. 50–57, 2004.
16. C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," In Privacy-preserving data mining, Springer, pp. 11-52, 2008.
17. E. Bertino, I. N. Fovino, and L. P. Provenza, "A framework for evaluating privacy preserving data mining algorithms*," Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 121–154, 2005.
18. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.
19. S. Wu and H. Wang, "Research on the privacy preserving algorithm of association rule mining in centralized database," in, 2008 International Symposiums on Information Processing (ISIP). IEEE, pp. 131–134, 2008.
20. A. Gkoulalas-Divanis and V. S. Verykios, "Association rule hiding for data mining," Springer Science and Business Media,vol.41, 2011.
21. A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern mathing in insuran claims using data mining tehniques," in 2017 Tenth International Conferene on Contemporary Computing (IC3). IEEE, pp. 1-7, 2017.
22. Breast cancer wisconsin (original) data set. (n.d.). ([Online] Available: https://archive.ics.uci.edu/ml/ machine-learning-databases/ breast –cancer -wisconsin /breast- cancer -wisconsin. data, [Accessed: May-2018]).