PRIVACY PRESERVING -DATA CLASSIFICATION METHOD ON VERTICALLY PARTITIONED DATABASE

Aju Mohan[#1], Nidhin Ravindran[*2]
[#1] Department of CSE, Gurudeva Institute Of Science & Technology ,Kottayam
APJ Abdul Kalam Technological University
Kerala, India
[1] ajumohan454@gmail.com
[2]nidhinravindran@gmail.com

***Abstract*—Data privacy is an important aspect of information sharing. The growing needs of multiple parties interaction in corporate and financial sector emphasize the need of developing privacy preserving and efficient distributed data mining algorithms. In this paper, proposes an associative classification model on vertically partitioned databases. Vertical partitioning takes the columns of a table in a database and places them in two or more other databases. Since the data comes from a single table and is contained in one location the resulting partition is more manageable than a situation in which several tables are in the same database file. To ensure data privacy, we design an efficient homomorphic encryption scheme and a secure comparison scheme. We then propose a cloud-aided frequent itemset mining solution, which is used to build an association rule mining solution.**

Keywords- Associative classification; Vertically partitioned data base;, Distributed data mining; Privacy preserving.

I INTRODUCTION

Distributed data mining [1] playing great role in extracting knowledge from geographically distributed sources. The exposures of distributed data mining include sensor networks, mobile ad-hoc communications, context aware computing, weather forecasting, intruder detection systems and web mining etc. The corporate and financial sectors as they spread over different geographic locations transforming their business intelligence applications to
distributed application form conventional centralized data
warehouse applications. Even though centralized data warehouse based models are more accurate than distributed applications, due to their easy scalability, communication efficiency and privacy facilities distributed data mining applications attracting lot of reasearcher's attention.

The works required in the privacy preserving data mining area are as follows:

- ➢ Privacy Preserving Data Publishing: These techniques try to study different techniques associated with privacy. These techniques consist of:
  - ▪ The Randomization Method: In this technique, any random value is added to the original value of the data to mask the values of the data. The noise is added in large amount so that the original data value is not recovered [3].
  - ▪ The K-Anonymity Model and L-Diversity: In K-Anonymity, the techniques like generalization and suppression were introduced to normalize data representation. In order to reduce the identification risk, every tuple in the database must be indistinguishable. The L-Diversity method was introduced to overcome some weaknesses of K-Anonymity. The new concept of intra group diversity of sensitive and private values within anonymization scheme was discovered [4].
  - ▪ Distributed Privacy Preserving: Sometimes, some users do not wish to disclose their information to other users. But the individual users are interested in achieving the aggregate results from the data set which are divided among the users. [5]

- ➢ Modifying the record values to preserve privacy: In this technique, Association Rule Hiding methods were used to preserve privacy. Using these methods, the association rules are encrypted in order to secure the data.

- Query Auditing: In this technique, either the result of the query is modified or the result of the query is restricted. Many Perturbation methods are used to achieve this. There is a lack of classification rules over the vertically partitioned databases,so we are proposing a method for classify the vertically partitioned database.

II. RELATED WORK

Fuzzy association rule based classification approach is proposed by Raghuram. Irrespective differences in association rule generation approach all these models implemented on horizontally distributed environment. In which training data set is horizontally partitioned among P processors and in every iteration each sites calculates local counts of candidate set and broadcasts these to all other processors where global classification rules will be generated. The limitations of these models are duplication of the entire set of candidates at each site due to which these models gives huge communication cost and multiple scans at each site will cause reduction in time efficiency. To the best of our knowledge there are no approaches present in the literature which can perform associative classification on vertically portioned databases.

Privacy-preserving algorithms on vertically partitioned data have been proposed with different techniques including association rules mining [2], [12] and classification [13],[14]. These privacy solutions can be broadly categorized into two approaches. One approach adopts cryptographic techniques to provide secure solutions in distributed settings. Another approach randomizes the original data in such a way that hides the underlying patterns [16] which can affect quality of results

III PROPOSED SYSTEM

In the proposed system, we follow some steps in the privacy preserving Associative classification model. In the distributed third-party setting as shown in "Fig. I" a database set D hold confidential databases Dl,D2 ..... Dn, respectively, each of which can be regarded as a relational table. Each database has same number rows. All this sub databases of D shares the sub set of variable of same transactions. There is a common ID that links the rows in distributed in among sites and all subset of transactions also holds the class label to which transaction belongs. A trusted third party data miner (DM) initiates the associative classification rules extraction process by broadcasting minimum support count. All sites generate transaction identifier (TID) list of each element separately for each class label. Using TID list all local sites generate frequent items of the particular c1asslabel and prepare attribute vector and scalar matrix of frequent items. The locally generated scalar matrix and private key encrypted attribute vector send to next site where scalar product will be performed to find global scalar matrix and vector. Since attribute vector is encrypted the successor site which received it can't predict the content. The next sites also repeat the same until all sites are covered. The final matrix and encrypted vector send to DM which holds public key of all sites to generate associative classification rules. The sites exchanging only scalar matrix and encrypted vector which ensures privacy.

The detailed steps for performing the protocol is as follows:

A. *Trusted third party based privacy protocol*

Step.1 he process will be started by DM by broadcasting MinSup threshold.

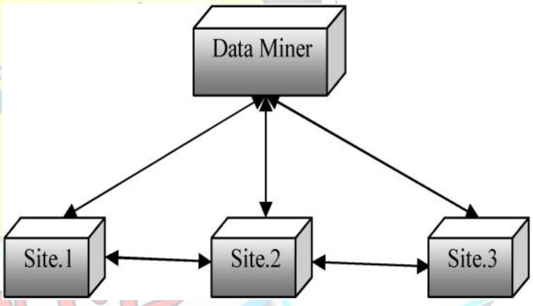

Fig.1 Communication between data miner and sites

Step 2. All sites forms TID list and finds local class label based frequent item sets.

Step3.Using frequent itemset obtained in previous step site prepares a scalar matrix Mi separately for each class label in which each row represents frequent itemset and column represent transaction.

Step4. Each site generates its item vector Vi separately for each class label which consists names of local frequent items as shown in section C. The vector Vi will be encrypted using private key and forms en Vi. The public key of site sends to DM.

Step5. The first Sitel sends matrix MI and the encrypted frequent item set list enVI to Site2.

Step6. The Site2 using M2 and V2 generated in step 2, 3 performs scalar product of M), M2 of same class label and using the frequent elements of scalar product matrix MI2 and enVI2 .

Step 7. Site2 then prepares a matrix M2' which consists of MI, M2 and MI2 and a vector enV2' which consists of encrypted frequent item set lists enV), enV2 and enVI2 separately for each class label. Site2 sends matrix M2' along with vector en V'2 to its successor site and its public key to DM.

Step8. Each Site i in the remaining sequence of Sites performs step 6 and 7 using received matrix and vector (M'i_l, enV'i_l) from its predecessor site and forms global matrix (M\) & vector (enV\).

Step9. Finally the list will reach to DM. The DM applies decryption algorithm using private keys received from sites for each element in the vector enV'n in corresponding order to get the frequent itemsets.

Step10. The decrypted frequent item sets are nothing but global frequent item sets. The DM prepares a list which consists of global frequent item sets with their support values separately for each class label.

Step11. Using global frequent item sets in the list, DM generates association rules of each class label as shown in section D.

Step12. The final rules are broadcasted to all sites and classifier implemented as stated in section E.

B. *Algorithm*:

Frequent item set generation in vertically fragmented databases. In order to generate class label based frequent item sets with single scan to database we proposed algorithm based on Fadi[6] Multi-class Classification model. The algorithm in its first scan itself generates Transaction id (TID) list of each element using which it calculate support of an itemset. In generating TID list the algorithm consider class labels of transactions also. Different TID list will be generated for different class labels and same elements with different class label will be considered as two different itemsets. The frequent item sets are those itemsets that satisfied the support threshold and only frequent itemsets used for generating new itemset. The algorithm will continue until no new item set can be generated. The detailed step of the algorithm as follows:

- *Stepl*. Scan the data base and form the TID list of each element separately for each class label by tanking elements as a row and all transactions as column. If the element present in a transaction mark intersecting field as 1 else mark filed as O.
- *Step2*. For each element in the TID list as item calculate their support count simply by counting number of 1 's present in the corresponding row.
- *Step3*. If support count of the item crosses the minimum support add it to frequent item list along with class label.
- *Step4*. Generate next level item sets using items of same class label in the frequent item list and calculate its support count from TID list simply by counting number of 1 's falling in same column of those items present in itemset.
- *Step5* . If itemset support count is greater than support threshold add it to frequent item list
- *Step6*. Repeat step 4 and 5 until no rule item pass support threshold.

IV CONCLUSION

Many of classification algorithms like decision trees are proposed on vertically partitioned data bases, but there is lack of associative classification on vertically partitioned data sources. In this paper, we proposed a third party based privacy preserving associative classifier for vertically partitioned databases. The experiment conducted on VCI data sets proved that our model shown good accuracy measures. We also showed analytically that our model is time, cost and privacy efficient.

REFERENCES

[I] M. Zaki, "Parallel and Distributed Data Mining: An Introduction," Large-Scale Parallel Data Mining, LNCS of Spinger, volume 1759,pages 1-23 ,2000.
[2] J. Vaidya, and C. Clifiton, "Privacy Preserving Association Rule
Mining in Vertically Partitioned Data," Proceedings of SIGKDD, Canada, 2002.
[3] B. Liu, W .. Hsu and Y.Ma," Integrating classification and association rulemining. Knowledge discovery and data mining" , Proceedings of Knowledge discovelY and Data mining, pp. 80--86, 1998.
[4] D.E. O'Leary," Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines," IEEE Expert, vol. 10, no. 2, pp. 48-52, 1995.
[5]W. Li, J. Han and J. Pei.. "CMAR: Accurate and efficient classification based on multiple class-association rules" ,Proceedings of ICDM, pp. 369-376,2001
[6] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. Proceedings of SIGMOD,2000

[7] F. Thabtah, P. Cowling and S. Hammoud , "MCAR: multi-class classification based on association rules," , Proceedings of the ACSIIEEE 2005 International Conference on Computer Systems and Applications, Washington, DC, USA, pp. 33-1. (2005)

[8] G.Thakur and C.J. Ramesh, "A Framework For Fast Classification Algorithms," International Journal lnformation Theories & Applications, V. I 5, pp. 363-369, 2008

[9] D. Mokeddem, H. Belbachir. "Distributed classification using class association rules mining algorithm," IEEE International coriference on Machine and web intelligence, Algeria,2010.

[10] B. RaghuRam and G. Aghila "Mobile agent based fuzzy associative classification rule generation for OLAM", IEEE International conference on lAMA, Chennai, India, pp 1-5,2009.

[11] B. RaghuRam and J. Gyani "Fuzzy Associative Classifier for Distributed Mining", proceedings of ICWET, Mumbai, India, pp 43 I -435, 2012.

[12]B. Rozeberg, and E. Gudes "Association rule mining in vertically partioned databases", Data and Knowledge Engineering, Elsever, pp378-396,2006.

[13] Z.Yang,and N. Wright "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol. I 8, No.9, pp 1253-I 265, 2006.

[14] H. Zheng, and S. R. Kulkarni "Attribute distributed learning: Models,limits, and aglorithms" .IEEE Transactions on Signal processing, VoI.59,no. I, pp386-398,2011.

[15] Y. Sang, H. Shen,H.Tain, "Privacy preserving tuple matching in distributed databases", IEEE Transactions on Knowledge and Data Engineering, Vol.2 I, No. 12, pp 1767-1782,2009.

[16] A. Delis and V.S. Verykos, and A.Tisonsis, "Data pertubaration approach to sensitive classification rule hiding,"Proceeding of ACM SAC 10, New York, USA, pp.605-609, 2010

[17] N.V Muthulakshmi and K. Sandyarani, "privacy preserving assoiciation rules mining in vertically partitioned databases". Proceedings ofIJCA,VoI.39, No. I 3 pp 29-35, 2012

[18] J. Natwich, X. Sun, and X. Li "Data reduction approach for sensitive associative classification rule hiding", Proceedings of ADC 08,ACM, Volume. 75, pp 23-30, 2008

[19] R. Agrawal,R. Srikant." Fast algorithms for mining association rules in large databases", in Proc. 20th Int, Con!, VLDB, pp. 478- 499,I 994.

[20] J. Quilnlan "C4.5: programs for machine learning," published by Morgan Kufman, 1993.