Privacy Clustered Mining of Association Rules in Distributed Database by Using Fuzzyfication

Shital K. Somawar, Prof.Vanita Babanne

M.E, Department of Computer Engineering, R. M. D. Sinhgad College of Engineering, Pune, India.

Professor, Department of Computer Engineering Department, R. M. D. Sinhgad College of Engineering, Pune, India.

ABSTRACT: Data Mining is very emerging technology for mining efficient dataset according to the user request. Association rules are very crucial part in data mining which tell the relationship between two item sets by using the apriori algorithm so that it will give the frequent item sets. Algorithm used in the system is EDMA (Efficient Data Mining Algorithm) which processes the user query and navigates to the appropriate databases which are local to that site. In First step apriori algorithm is used to find the association of requested dataset and in second step the fuzzy association rules are taken place by applying the fuzzy operations. The system uses the fuzzy association so that if any item does not get the association with any other item then it will be associated because of the complex fuzzy operations. So that it will be better to get the requested data to the user because the system will check the datasets on each and every site and by combining the datasets of different sites it will be given to the user. Thus the transaction will be secured by using the association key and also user requested data will be given and also the system is used as a analytical tool which provides the precise data.

KEYWORDS: Data mining, Association Rule, Fuzzy operations, Fuzzy Association Rule Mining, Frequent item sets, Privacy-preserving.

I. INTRODUCTION

Now a days data is very crucial part for any process and for getting the data the various methods take place. And one of the approaches for getting the data is data mining approach. Mining of data give the related information regarding specific subject. Data mining is emerging technology for mining efficient and effective datasets according to the user request. For mining the requested datasets according to the user request the association rules are used. The centralized data is very difficult to handle and load balancing is very hard to maintain. The problem of load balancing and maintenance is resolved by distribution of data among several sites. The Extraction process of association rules securely from distributed databases is very difficult. Some of the extraction techniques do not mine the data securely so that attacker can easily get the data like military data, bank data or any essential data. Some mining techniques are very expensive in terms of computational and communication cost. Data is transmitted among the several sites as the privacy is the main issue. There are many protocols and methods take place to preserve the privacy of the datasets like clustering, classification, association rule mining, decision tree, bagging, boosting, neural networks, combined mining.

There are articles addressing various problems such as the collusion free communication and privacy preserving of datasets which reside on different sites were resolved by using various techniques and methods of data mining. Our proposed system focuses on the privacy as well as the user requested dataset by using the fuzzy association rules. Every item which is reside on the different sites, which firmly extracts the association rules which is relevant to the users requested dataset and has the communication and computational cost is less than formerly used techniques.

There are ample of fuzzy operators can be designed to comply the crisp set theory all the alternatives of AND, OR and NOT operators behave exactly the same for the combinations on membership. The behavior of fuzzy operations is same as that of the Boolean operators. There are several types of fuzzy operations like simple, complex and medium for different behavior of data. According to the calculated support value the operations are applied.

II.RELATED WORK

R.Agrawal and R.Srikant et al. [1] fast distributed mining algorithms to improve the experimental results.Paper focuses on association rules between items of large databases by using two new algorithms. They have combined the features of two algorithms into hybrid algorithm to scale up the number of transactions.
Vassilios S. Verykios, Elisa Bertino et al. [2] privacy preserving data mining. They have proposed the concept of classification hierarchy and coordinate work on classification hierarchy and evaluation of performance.
Murat Kantarcioglu,Chris Clifton et al.[3] which describe privacy preserving distributed mining of association rules in horizontally distributed databases. This paper addresses secure mining of association rules over horizontally partitioned data. The methods integrate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task.

Dragos Trinca and Sanguthevar Rajasekaran et al.[4] describes the collusion free mining of association rules in vertically distributed database and also provides the privacy by using the multiparty protocol. This paper focuses on the case when the database is distributed vertically, and proposes an efficient multi-party protocol for evaluating item sets that preserves the privacy of the individual parties. And it is much faster and secure.
JIANG Dongjie,XUE Anrong et al.[5] describes the privacy preserving database which is horizontally divided. This paper focuses on privacy preserving data in semi-honest circumstances and also uses protocol such as the secure sum computation, scalar product computation, standardization and computation by means of semi-honest third party and algorithm resolves the problem of privacy preserving of data in semi-honest environment.

Madhuri N.K.,Reena K. et al.[6] describes the privacy preserving association rule mining algorithm which satisfies the privacy constraints. This paper focuses on different methods for mining association rules.
JayantiDansana,RaghvendraKumar,Debadutta Dey et al.[7] Association rule mining finds the strong association between the large items of data sets. This paper mainly focuses on the privacy using secure sum in randomized response technique. Because when the data is distributed among different sites in horizontal partitioning across multiple sites and the sites wish to collaborate to identify valid association rule globally.

Yiqun Huang, Zhengding Lu, Heping Hu et al.[8]they have explain the privacy preserving data mining process takes place of vertically partitioned data. Secure multiparty computation takes place by using the scalar product.
Alex Gurevich, Ehud Gudes et al.[9]they focus on three algorithms for finding association rules.
Jie Liu, Xiufeng Piao, and Shaobin Huang et al.[10]This paper addresses secure mining of association rules which builds the hash table to prune the item sets and cryptographic techniques for security and privacy purpose.

III. IMPLEMENTATION DETAILS

A. *UNIFI-KC Protocol:*

Phase1:Each player selects cipher and encryption key and depending on that player calculates the hash function to build the lookup table.
Phase2: Fuzzy association takes place according to the requested data by predicting with other sites and support and confidence is calculated using the efficient mining algorithm.
Phase3:Encryption of all items sets using the hash function.
Phase4:Merging of all item sets where each odd player sends the item sets to p1 and each even player sends the item sets to p2 and finally p2 sends all item sets to p1.
Phase5:Decryption of item sets at the user level by using the each players private key.

B.. *EDMA Algorithm:*

1.Create Main Site and N Computing Site
2.Task of Main Site Generation of 1-item frequent set
3.Divide the data set in N parts and send it to all computing Sites
4.Wait for receiving 1-item Candidate sets as messages from each computing Site
5.Combine all the small candidate sets to form a new candidate Set- >1-Candidate

6.Call Genfreqset(1-Candidate)

7.Send 1-Frequent set as message to each computing Site for finding small 2-item candidate sets.

Genfreqset(k-Candidate)

8.If k-Candidate is empty or k equals number of items in the data set **-** > terminate

9.k-Frequent< **-**prune k-Candidate using support threshold

10.Store and return the k-Frequent set Collectcandset( )

11.Wait for receiving k-item Candidate set as messages from each computing Site

12.Combine all the small candidate sets to form a new candidate set**-** >k-Candidate

13.Call Genfreqset(k-Candidate)

14.If Genfreqset terminates

15. Fuzzy association rule

16.Getting Frequent item set Union or intersection are primary operation are use

      IfSize(Local(itemset))=size(distributed(itemset))

      Then use primary fuzzy operation

      IfSize(Local(itemset))>size(distributed(itemset))

      Then use Complex fuzzy operation

      IfSize(Local(itemset))<size(distributed(itemset))

      Then use Semi Complex fuzzy operation

17.Combine all the Frequent itemsets and return it

18.Send terminate as message to each computing Site

19. terminate

20.Else send the k-Frequent set as message to each computing Site for finding small (k+1)-item candidate sets Call Collectcandset( )

IV.SYSTEM ARCHITECTURE

The system architecture describes the how system actually works. There are several sites and each site has its own user having its data and trying to request the data as per his interest to particular site.

Data is encrypted using sites key and according to the requested data and sites available data the association rules are generated using particular fuzzy operations. Relative to the particular support value the operations are applied like simple, complex and medium type. The pattern of data is determined. And at last decryption of data takes place and data is provided.
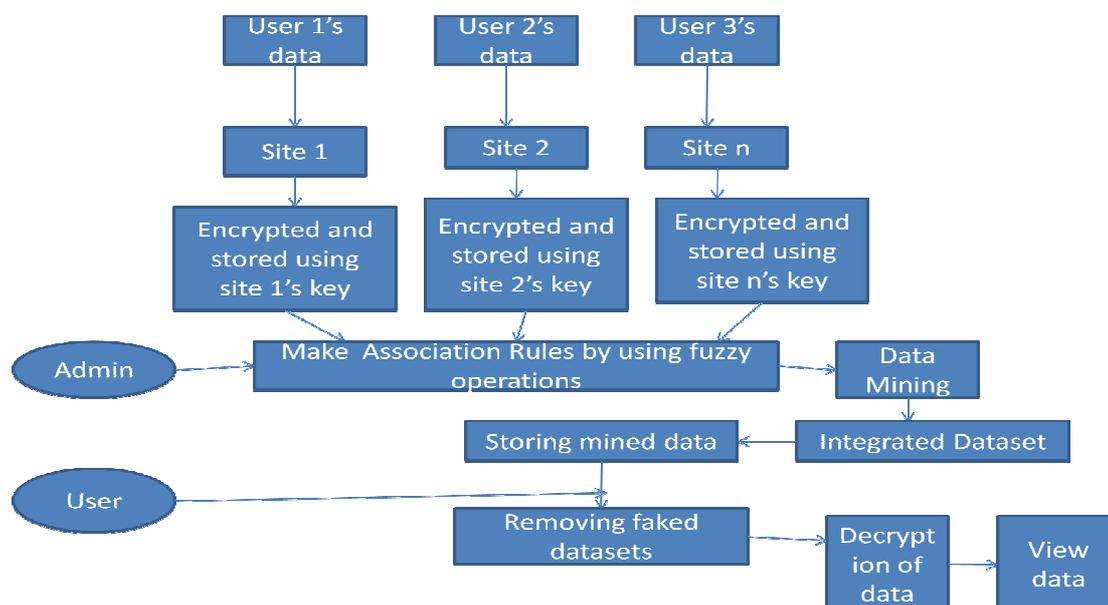

Figure 1.System Architecture

V.MATHEMATICAL MODEL

A.SET THEORY:

  Consider a function DM {D, S,N,AD} It describes the mathematical terms used in the system.

  1.User requested datasets which are in terms the characters and distributed among the several sites.
$$Input\ dataset\ D = \{d1,d2,…,dn\}$$
  2.Sites which are distributed and each site contain its own local dataset.
$$Sites\ S=\{s1,s2,…, sn\}$$
  3.Users of the systems who request the particular datasets and gets the datasets in the encrypted format and   they will decrypt that dataset by using the their private key.
$$User\ N=\{n1, n2,…, nm\}$$
   4.Output Datasets which contains the precise dataset as per the user query and provide to the user in encrypted format.
$$output\ Datasets\ AD =\{ad1, ad2,…, and\}$$

  B.Mapping Diagrams:
In mapping diagram it describes the mapping relationships such as one to one, many to one and many to many. The below figure describes the mapping of datasets to the different sites.It shows that same data can be present on different sites. Means data is distributed among the different sites.
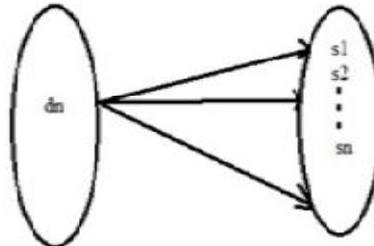

Fig. 2. Mapping Of Datasets

The second diagram shows the mapping between the global database and different sites. Global datasets can have multiple number of sites which has their own database.
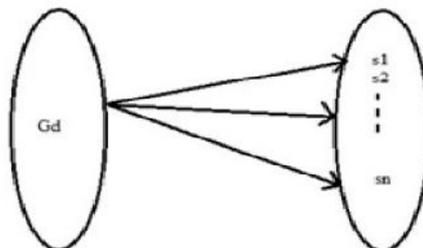

Fig. 3. Mapping Of Global Datasets

VI.RESULT

Dataset:For implementation any groceries dataset D is used. So that by using that association rules are generated as this is very fundamental part of data mining. System shows the expected results which provide the user data in which the user is interested and in minimum time. It takes the minimum computation time with use of fuzzy operations. Consider Apriori computation using fuzzy with FDM

S:threshold value
M:Total players
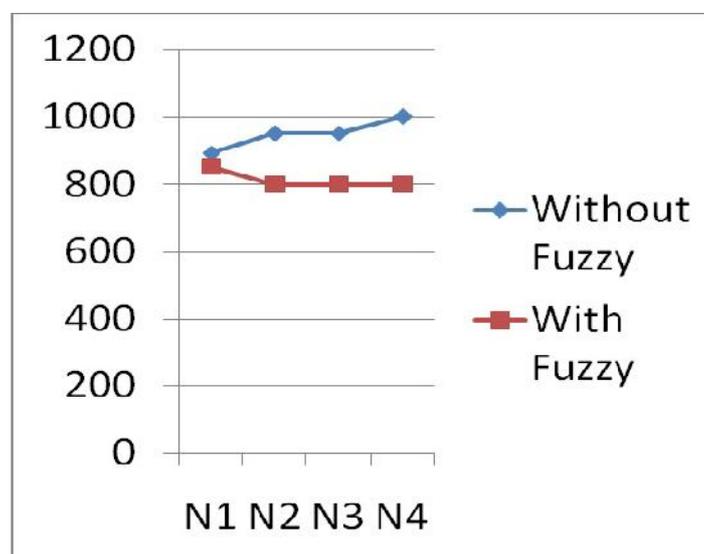N.Experiment set
N=200,400,600,800,1000
M=4

Figure 4:Graph

V.CONCLUSION

     EDMA provides an efficient method for generating association rules from different datasets, distributed among various sites. Our proposed optimization technique increase linearly as we increase the number of sites. It exchanges fewer messages than FDM using fuzzy set given by k-mean, so EDMA is more scalable. On the other hand, candidate set broadcasts its candidate support counts to all other sites and subsequently receives support counts of others, broadcasting numerous messages when we increase number of sites.ARM (Association rule mining) perform efficiently on different organizations in different domains. As security and privacy is a common issue for data mining application,DARM (Distributed Association rule mining) maintains the privacy without increasing overall communication costs.

REFERENCES

[1] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proc 20th Intl Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.
[2] Vassilios S. Verykios, Elisa Bertino, State-of-the-art in Privacy Preserving Data Mining, SIGMOD Record, Vol. 33, No. 1, March 2004.
[3] Murat Kantarcioglu,Chris Clifton, Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 9, SEPTEMBER 2004.
[4] Dragos Trinca and Sanguthevar Rajasekaran,Towards a Collusion- Resistant Algebraic Multi-Party Protocol for Privacy-Preserving Association Rule Mining in Vertically Partitioned Data, 2007 IEEE.
[5] JIANG Dongjie,XUE Anrong, Privacy-preserving DBSCAN on Horizontally Partitioned Data, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education.
[6] Madhuri N.K.,Reena K., Privacy Preserving Mining of Association Rules on Horizontally and Vertically Partitioned Data, 2012 IEEE.

[7] Jayanti Dansana,Raghvendra Kumar,Debadutta Dey, Privacy preservation in using horizontally partitioned databases using randomized response technique, Proceedings of 2013 IEEE Conference on Information and Communication Technologies.

[8] Yiqun Huang, Zhengding Lu, Heping Hu, A Method of Security Improvement for Privacy Preserving Association Rule Mining over Vertically Partitioned Data IEEE 2005.

[9] Alex Gurevich, Ehud Gudes, Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation, IEEE 2006.

[10] Jie Liu, Xiufeng Piao, and Shaobin Huang, A Privacy-preserving Mining Algorithm of Association Rules in Distributed Databases, IEEE 2006