International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 7, July - 2014

A Secure Approach in Getting Association Rules from a Service Provider for Privacy Preserving in Market Basket Database

Prakash Soma
M.Tech(ComputerScience)
Department of Computer Science
School of IT,JNTU-Hyderabad
Telangana, India

Uma Rani.V
Assistant Professor
Department of Computer Science
School of IT,JNTU-Hyderabad
Telangana, India

***Abstract*—Data mining provides useful information such as frequent patterns and association rules for decision making in marketing development strategies. Now a days, many software companies are offering data mining services to various customers. At the same time most of the customers like small and medium size store's owners are not able to perform their data mining tasks on their own. These customers are willing to get data mining services from service providers. Sharing sales information with service providers involves losing privacy of the customers' personal information and valuable business details .In this paper we discusses possible security and privacy threats in getting association rules from service providers and suitable solution to it.**

Keywords: Data mining, Decision making, service providers, security and privacy threats, association rules

I. INTRODUCTION

Data mining allows us to find useful information such as association rules, frequent patterns from a large database. Association rules are one of the representation of the mining result. In retail industry, the database consist data on sales, customer purchasing history, goods transportation, consumption and services. Mining information in market basket database provides the customer buying patterns and trends which can be used to improve customer service quality, good customer retention, store alignment strategy. Data mining resulting information also useful in financial data analysis, retail industry, telecommunication industry, biological data analysis, intrusion detection and so on. So most of the small and medium size retail industries are willing to mine their store's databases. As they are not capable of having technology and man power, they want to get association rules from service providers. Many software companies are providing data mining services to their customers.

II. PROBLEM DEFINITION

A. Security threats

When a retail shop owner wants data mining service from a particular service provider, located geographically at remote location, it involves various security and privacy issues.

Confidentiality and Integrity

As the service provider located at different location ,the customer must send his input database files through network. At the same time, the service provider must send his mining result i.e. association rules through network only. There might be a number of security attacks on database files and mining result in network. Intruders may disclose the data base file's information and mining result or modify the input data and mining result. The confidentiality and integrity of the files that are going to be sent over the network will be in doubt.

B. Threats to the input data privacy

All service providers are may not be trust worthy. Some service providers may keep input data provided by the retail shop owner and result of the mining result with them and may be used by them in future for their business purpose. They may also sell our valuable mining results to our market competitors. So, The privacy of input data and mining results in dilemma. This is a serious drawback in getting mining result from service providers.

III. PROPOSED METHOD

In the proposed method various methods are used to resolve the security and privacy threats to the input database and mining results. It consist encryption and decryption methods to provide confidentiality, message digest algorithm is used to provide integrity services and data transformation, mapping techniques for privacy preserving service to the input database .

The basic block diagram of proposed architecture is as shown below.

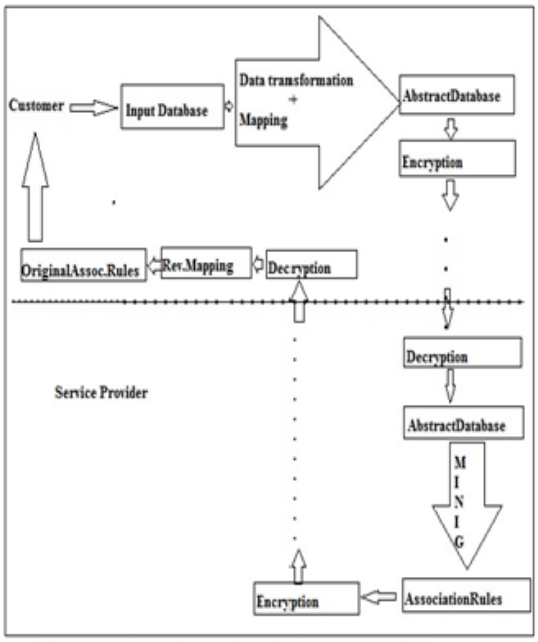

Fig.1.Basic block diagram of proposed architecture

A. Data transformation & Mapping for privacy preserving

Data Transformation techniques provides a statistical guarantee of data confidentiality. The main goal of transformation is to make irreversible data modification and destroy actual values and correlations among them. The main idea in data transformation is to preserve aggregate trends in original data while changing the original data. For example, data may be interchanged between different rows to hide exact mapping between fields of a given record, noise data may be added to the data up to some limit. If store owner sends original sales database to service providers, there may be a chance of misusing the customers' personal details and sales information of the store. Sales information may be shared with business competitors. To preserve privacy in the database, we can use various mapping and replacement techniques. Some of them are mentioned below.

The below table consist 5 transactions each having different items.

TABLE I: GENERAL REPRESENTATION OF SALES DATA

Transaction	Items
T1	Bread, Milk, Sugar, Rice
T2	Milk, Sugar, Rice, Salt
T3	Bread, Rice, Salt
T4	Bread, Milk, Sugar
T5	Milk, Rice

In the above general retail database the items in transactions are represented with original items' names but we can use generic symbols like True/False,0/1 or Yes/No instead of using exact names of items.

TABLE II: ALTERNATIVE REPRESENTATION OF SALES DATA

	Bread	Milk	Sugar	Rice	Salt
T1	*1*	1	1	1	0
T2	*0*	1	1	1	1
T3	*1*	0	0	1	1
T4	*1*	1	1	0	0
T5	*0*	1	0	1	0

Items' names can be mapped with some generic variables so that service provider cannot understand the association among the items.

TABLE III:MAPPING BETWEEN ITEMS AND SYMBOLS

Items	Symbols
Bread	A
Milk	B
Sugar	C
Rice	D
Salt	E

After performing mapping and replacement activities the sales database that is to be sent to the service providers is to be as below.

TABLE IV: ABSTRACTIVE REPRESENTATION OF SALES DATA

	A	B	C	D	E
T1	*1*	1	1	1	0
T2	*0*	1	1	1	1
T3	*1*	0	0	1	1
T4	*1*	1	1	0	0
T5	*0*	1	0	1	0

By sending only generalized sales database to service providers we can preserve privacy of customer's personal information and secrecy of valuable business transaction database.

B. Using MD5 to provide integrity

MD5 algorithm was developed by Ronald L.Rivest. It is mainly used for digital signature applications, in which a large file must be encrypted with a secret key.

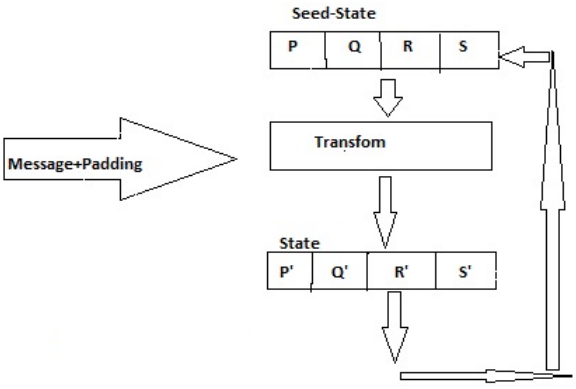

Fig.II. Basic block diagram of MD5 algorithm

MD5 algorithm takes arbitrary length message as input. Input messaged is padded so that its length in bits equals to 448 mod 512.Padding is always performed though the message length more than 448 bits. A four word buffer(P,Q,R,S each 32 bit register) is used to compute the message digest. These registers are initialized with the following values in hexadecimal, lower order bytes first.

Word P:01 23 45 67
Word Q:89 ab c def
Word R:fe dc ba 98
Word S:76 54 32 10

Output of the MD5 algorithm is to be appended to the input database.MD5 algorithm is more concerned with security than speed. It will provide the integrity service to the input data at receiver side.

C. Using DES to provide confidentiality

DES(Data Encryption Standard) is a symmetric block cipher encryption algorithm with 64 bit input text blocks and 56 bit key. The input to this DES algorithm is the database along with the MD5 message digest. This algorithm generate 64-bit output block for each 64-bit input block. Initial key is used to generate 16 sub keys for 16 rounds. The basic structure of DES algorithm is as shown below.

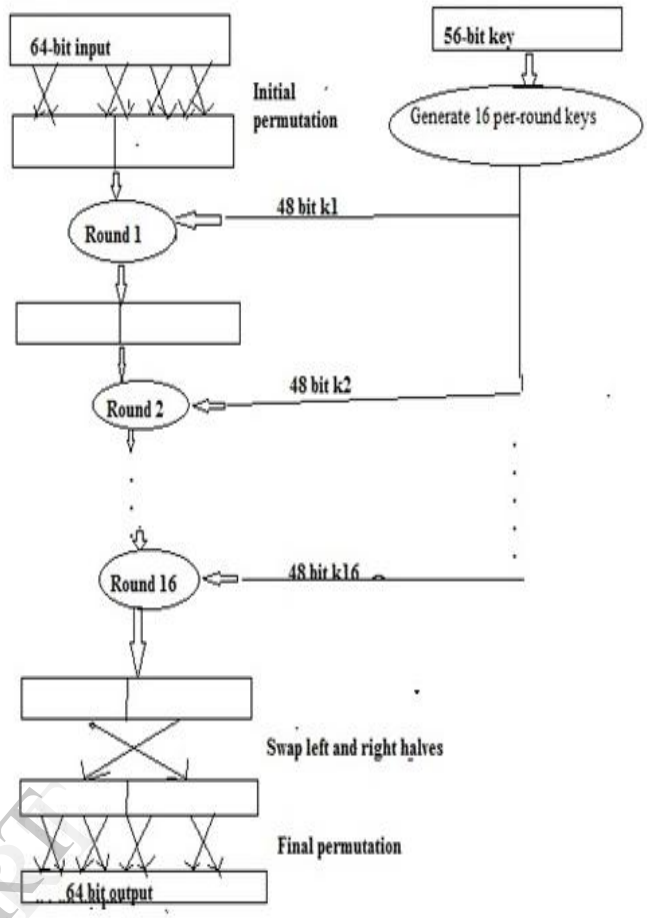

Fig.III. Basic block diagram of DES Encryption algorithm

The advantages of using DES algorithm are

- It provides confidentiality.
- It is relatively fast encryption algorithm.
- Keys generated in this algorithm are structurally sound.

This encrypted data is to be sent to the service provider. The service provider decrypt the encrypted data and perform mining operations. Sending of the mining result from service provider to customer is same as the above procedure.

D. Association Rules generation using Apriori Algorithm

The service provider uses Apriori algorithm to generate association rules. Apriori algorithm was proposed by R.Agrawal and R.Srikant in 1994 for mining frequent itemsets for Boolean association rules. This algorithm uses prior knowledge of frequent itemset properties. It employs an iterative approach known as a level-wise search, where k-itemsets are used to find (k+1)-itemsets. First, the set of frequent 1-itemsets is found by the entire database to accumulate the count for each item, and collecting those items that satisfy minimum support. L(1) is considered as resulting set. L(2),the set of 2-itemsets is to be found by using the L(1).L(2) is used to find L(3),and so on , until no more frequent k-itemsets can be found. One full scan of the database is required to find each L(k). To improve the efficiency of level-wise generation of frequent itemset, Apriori property i.e. “All nonempty subsets of a frequent

itemset must also be frequent" is used. The steps in Apriori algorithm to get association rules are shown below

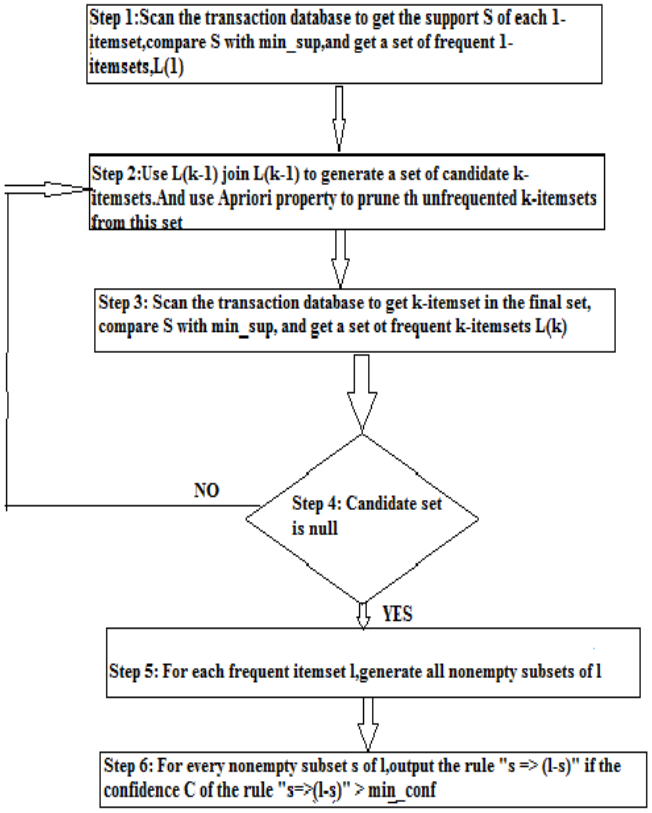

Fig. IV. Steps in Apriori algorithm to get association rules

Association rules are generated using the frequent itemsets of Apriori algorithm .Strong association rules are association rules which satisfy both minimum support and minimum confidence. Confidence of an association rule can be calculated as follow:

$$\text{Confidence}(X \Rightarrow Y) = P(Y/X) = \frac{\text{support\_count}(XUY)}{\text{support\_count}(X)} \quad (1)$$

The conditional probability is expressed in terms of itemset support count, where support_count(XUY) is the number of transactions containing the itemsets XUY, and support_count(X) is the number of transactions containing the itemset X. Using confidence association rules are generated as follows:

- For each frequent item l, generate all nonempty subsets of l.
- For every nonempty subset s of l,output the rule "s=>(l-s)" If support_count(l)/support_count(s)>= minimum confidence threshold.

As rules are generated from frequent itemsets, each association rule satisfies minimum support.

E. Finding Sensitive Items by customer

After receiving encrypted abstractive association rules from service provider, the customer decrypt and remaps them. After getting original association rules, customer selects frequent item's single antecedent association rules with higher confidence. Then he combines all the consequents of the same single antecedent item. Then right side of the items in the combined single antecedent association rules are taken as sensitive items. These sensitive items are useful for hiding sensitive association rules in the sales database to provide privacy to the input database before it is published.

IV.EXPERIMENTAL RESULT

The output of our of experiment with following input is as follows:
Input:
Database : TABLE IV
Minimum Support= 2
Minimum Confidence: 0.40

Output:

Frequent Itemsets generated from given database using Apriori algorithm:
Frequent 1-Itemset:
{A,B,C,D,E}
Frequent 2-Itemsets:
{{A,B},{A,C},{A,D},{B,C},{B,D},{C,D},{D,E}}
Frequent 3-Itemsets:
{{A,B,C},{B,C,D}}

Association Rules generated from frequent item sets:
A ==> D #SUP: 2 #CONF: 0.666
D ==> A #SUP: 2 #CONF: 0.5
B ==> C #SUP: 3 #CONF: 0.75
C ==> B #SUP: 3 #CONF: 1.0
B ==> D #SUP: 3 #CONF: 0.75
D ==> B #SUP: 3 #CONF: 0.75
D ==> E #SUP: 2 #CONF: 0.5
E ==> D #SUP: 2 #CONF: 1.0
A B ==> C #SUP: 2 #CONF: 1.0
B C ==> A #SUP: 2 #CONF: 0.666
A C ==> B #SUP: 2 #CONF: 1.0
C ==> A B #SUP: 2 #CONF: 0.666
B ==> A C #SUP: 2 #CONF: 0.5
A ==> B C #SUP: 2 #CONF: 0.666
B C ==> D #SUP: 2 #CONF: 0.666
C D ==> B #SUP: 2 #CONF: 1.0
B D ==> C #SUP: 2 #CONF: 0.666
D ==> B C #SUP: 2 #CONF: 0.5
C ==> B D #SUP: 2 #CONF: 0.666
B ==> C D #SUP: 2 #CONF: 0.5

After remapping the customer will find sensitive items according to his requirements manually.

V. CONCLUSION AND FUTURE WORK

In this paper, we discussed various security problems and privacy issues in getting association rules from service providers and also proposed some suitable solutions to them.

This procedure is suitable to small size of databases. If the database is very large then the performance issues of different algorithms should be evaluated and replaced with some suitable algorithms if necessary.

REFERENCES

[1] Jiawei Han& Micheline Kamber&Jian Pei(2011). Data Mining Concept and Techniques.
[2] William Stallings. Network security essentials applications and standards(4th edition).
[3] S.Kasturi & T.Meyyappan(2013)Detection of sensitive items in market basket database using association rule mining for privacy preserving, Proceedings of the 2013 international conference in Feb 21-22.
[4] Agrawal R, Imielnski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC.
[5] Agrawal R, Srikant R (2000) Privacy preserving data mining. In ACMSIGMOD Conference on Management Of a,Dallas,Texas, pp 439–4501.
[6] Clifton C, Marks D (1996) Security and privacy implications of data mining. In: SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery.
[7] Agrawal R, Imielinski T, Swami AN. "Mining Association Rules between Sets of Items in Large Databases." SIGMOD. June 1993, 22(2):207-16.
[8] Agrawal R, Srikant R. "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99,ISBN 1-55860-153-8.
[9] Mannila H, Toivonen H, Verkamo AI. "Efficient algorithms for discovering association rules." AAAI Workshop on Knowledge Discovery in Databases (SIGKDD). July 1994, Seattle, 181-92.
[10] Charlie Kaufman,Radia Perlman,Mike Speciner.Prentice-Hall India. Network Security PRIVATE communincation in a PUBLIC world.SecondEdition. .