International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 18
December 2016

Privacy Preservation: Chi Square Computation for Association Rule Mining

Raghvendra Kumar[1], Dr. Prasant Kumar Pattnaik[3], Dr. Yogesh Sharma[3]

[1,] *Ph.D Scholor, Jodhpur National University, Jodhpur, Rajsthan, India*
[2] *School of Computer Engineering, KIIT University,Bhubaneswar, India,*
[3,]*Jodhpur National University, Jodhpur, Rajsthan, India*
raghvendraagrawal7@gmail.com, patnaikprasant@gmail.com, yogeshsharma@gmail.com

Abstract

The big trouble of privacy preservation in data mining techniques has been studied comprehensively in current passing years because of the improved amount of private information which are presents locally and globally in the distributed database environments. The high amount of privacy preservation transformations use some form of data perturbation or representational ambiguity in order to reduce the risk of identification and security as well as privacy. In this paper, we propose an algorithm that is applicable all partitioned database may be horizontal, vertical and hybrid partitioning. The proposed algorithm provides the highest privacy preservation to the all database, because we used the data modification concept with the help of Chi Square concept and after that we used the privacy preservation algorithm to provide the privacy to the distributed homogeneous database by selecting the random number by all parties, then calculate the global support by using the algorithm that mentioned in the paper, with high privacy and zero percentage of data leakage.

Keywords: Data Mining, Distributed database, Association rule mining, Chi Square, Privacy preserving techniques.

1. Introduction

In the present years an increasing amount of individual data is being stored by confidential locally and globally distributed database. Since of advances in storage space in both hardware technology as well as software technology. This has main concerns about the both the security and privacy in the distributed data. The group of privacy preservation data mining is designed to develop tools, which can mine such sensitive data without compromising their privacy with the rest of the world. A number of models have been implemented by different researchers or authors to provide a different representation of the underlying data. A related field to privacy is that of uncertain data management. In uncertain data management, a variety of management and mining tools need to be implements on data which is specified only roughly. Typical models in uncertain data mining [1] assume that a probability distribution function of the data is known. This probability distribution is then used in order to construct and improve the effectiveness of data management models. There are two fields of privacy preservation modelling and uncertain data mining are closely related to each other, the research in the two fields has been largely independent. While privacy preservation transformations such as hash based techniques, randomization, cryptography based techniques used the probabilistic models, the end result cannot be effectively leveraged with uncertain techniques of data analysis. This is because such methods typically do not calibrate the added noise so that individual records can be used effectively with data management applications. In such cases, data mining techniques such as classification need to be Re-designed to work with aggregate distributions rather than individual records. In this paper we studied about the business data that is used the randomization techniques, chi square and Apriori algorithm. Mining association rules in transaction databases has been demonstrated to be useful and technically feasible in several application areas, particularly in retail sales.

1.1 Association Rule Mining

Association rule mining find the association or correlation among the large data items. Association rule shows attributes value condition that occurs frequently in a given data sets. Let I = {I_1, I_2. . . I_m} be a group of items. Let D be a set of transactions, where each transaction T is a set of items, such that T ⊆ I. An association rule is an implication of the form X ⇒ Y, where X ⊂ I, Y ⊂ I, and X ∩ Y = ∅. The association rule X ⇒Y holds in the database D with confidence C if C% of transactions in D that contain X also contain Y. The rule X⇒Y has support S if S% of transactions in D contain X ∪ Y. Mining association rules is to find all association rules that have support and confidence greater than user-specified minimum support and minimum confidence for example, beer and disposable diapers are items such that {beer →diapers} is an association rule mined from the database if the co-occurrence rate of beer and disposable diapers is higher than min sup and the occurrence rate of diapers in the transactions containing beer is higher than min conf. The first step in the discovery of association rules is to find each set of items that have co-occurrence rate above the minimum support. An item set with at least the minimum support is called a large item set or a frequent item set. In this paper, the term frequent item set will be used. The size of an item set represents the number of items contained in the item set, and an item set containing k items will be called a k item set. For finding the association rules in a given data sets we used the Apriori algorithm.

Apriori Algorithm1

```
L1 = {large 1-itemsets};
for ( k = 2; Lk-1 □ Ø; k++ ) do begin
    Ck = apriori-gen(Lk-1); // New candidates
    For all transactions t □ D do begin
        Ct = subset(Ck, t); Candidates contained in t
        For all candidates c □ Ct do
            c.count++;
    End
    Lk = {c □ Ck | c.count □ minsup}
End
Answer = □ k Lk;
```

R. Agrawal et. al [1] presented a first algorithm AIS for association rule mining. It was efficient algorithm for that time.

R. Agrawal and R shrikant [2] presented two new algorithms Apriori and Apriori TID that different from the previous proposed algorithm. They presented the experimental result using the both synthetic and real life data. The result of this algorithm is very outperformed than the previous algorithm. They also combined the best features of both these algorithm into the new algorithm is called Apriori Hybrid. This Algorithm has excellent scale up properties. It is also opened up the feasibility of mining the association rule over the large database.

U. Fayyad et. al [3] presented a framework for describing the link between the data mining, knowledge discovery and other related field of data mining. It was one of the initial efforts to relate the data mining with KDD.

T. Shintani et. al [4] presented four hash based parallel algorithms for association rule mining. Although these algorithm are good but actually these were developed to processed on parallel on processor.

E. Cohen et. al [5] presented an algorithm to find the interesting association without the support pruning. They employed this algorithm with combination of sampling and hashing techniques.

G.I. Web et. al [6] tried to find the association rule through the direct searching instead of two stage of processes of apriori algorithm. Author argue that apriori algorithm impose a large computational overhead when the database size is large.

J. Hipp et. al [7] presented a general survey and comparison of the different association rule mining algorithm.

G.K. Phalshikar et. al [8] proposed an algorithm for association rule mining using the concept of heavy item sets.

Y Xia et. al [9] proposed a algorithm for generalized the privacy preserving association rule mining problem by allowing the different attributes having the different level of privacy, that is using the different number of randomization in process. Author also proposed the efficient recursive estimation algorithm, to estimate the support of item set under this framework.

P. Wang [10] conducted a general survey of the different privacy preserving research and author also suggested some of the future direction, how to preserve the datasets.

1.2 Horizontal Partitioned Vs Vertical Partitioned Data

Horizontal (Homogeneous) partitioned data [11] [12], where each party collect data about the same attributes of object. For example, hospital that collect similar data about daises but for different patients. Vertical (Heterogeneous) partitioned database [11] [12], each party collects different attributes for same objects. For example, patients have attributes for a hospital that are different from attributes with Insurances Company. If the data is distributed vertically, then unique attributes that appears in a pattern or a transaction can be linked to the owner.

1.3 Privacy Preservation in Distributed Data Mining

Distributed data mining model assumes that the data sources are distributed across the multiple sites. The main challenge come here is to how to mine the data across the distributed sources securely without the either party disclosing its data to the other sites. Many algorithm are developed in the field of data mining to preserve the privacy [13] but they do not take into account because the focus only on the efficiency. A very simple approach to mining the private data over the multiple sources is to run existing data mining tool at each sites and combine the result. However this approach failed to give up the valid result for following main reasons.

1. Value for single entity may be spilt across sources. Data mining for individual sites will be unable to detect the cross site correlations.
2. The same item may be duplicated at different sites and will be over weight in results.
3. Data at single site likely to be from the homogenous population, important geographic and

demographic distinction between that population and others cannot be seen into the single sites.

1.4 Secure Multi Party Computation

The main problem of the parties that they perform the computation on the union of their private data but they don’t want to trust to each others, and each party want to hide their private data to the others, its refers to as the secure multi party computation. There have many problem and solution to perform the secure multi party computation [14] and solve the problem. Most of the solution assumes that one of the parties is trusted party and that party do not compute or distribute the results to the other party. In secure multi party computation contain two types of models first is Real model and second is Ideal model. Real model in which there are no any occurrences of the third party but in the ideal model there is chance of third party or trusted third party.

1.5 Chi Square Computation

Chi square is a classical techniques for measuring the closeness of two probability distribution, continues one of them is widely used for statically significance in many scientific circles e.g. biological etc. Chi square analysis is to be used to access the statically significance level of dependencies between antecedent and coincident of the rules. Chi square is defined in term of entries of the observed contingency table and expected contingency table as follows.

$$\chi^2 = \sum_{0\leq i,j\leq 1} \frac{(observed_{i,j} - expected_{i,j})^2}{expected_{i,j}}$$

Thus $\chi2$ represent a summed normalized square deviation of the observed values from the corresponding expected values

S. Brin, R. Mot-wani, and C. Silverstein [15] address the use of the chi squared significance level in the mining process itself, while we focus on computing the chi squared level independently of mining.

B. Liu, W. Hsu, Y. Ma [16] addresses the use of chi squared analysis for pruning the rule set and finding the set of representative rule after mining and appears to rely two dimensional contingency table.

R.J. Bayardo and R. Agrawal [17] use an expression for chi squared as convex function for support and confidence.

2. Problem Definition

Lets N number of parties {$P=P_1, P_2....P_n$} in the distributed database environments where N≥3. Each party have their own local database DB_i. We assume that parties are using the same protocol. The main goal to find the global support value that is applicable to all the local item sets and hides their private information or data to other parties presents in the distributed environments. In our algorithm we used the concept of chi square and random number to hide the private information or data from all other parties, when the number of parties are greater than or equal to three and each of the parties want to their global result without disclosing their local result.

3. Objective

The Objective of this paper is as follows

1. Each Party can identify only their own data

2. No any other party are able to learn their links between other parties and their data
3. No party learns any transactions of other party database
4. Provide the high privacy to their local datasets and compute the global result by using the Chi Square and Random number concepts.

4. Literature Survey

We studies different number of paper to fulfill our objectives

Lindell Y. and Pinkas B. [18] uses cryptography techniques as privacy preserving techniques in data mining. Through these techniques sensitive data can be encrypted, there is also proper toolset for algorithm of cryptography.

J. Vaidya and C. Clifton [19] proposed a privacy preserving techniques for association rule mining in distributed database into vertical segments.

Kargupta H. et.al [20] they tries to preserve the privacy of the data by adding random noise, while making sure that the random noise still preserve the signals from the data so that the patterns can still be accurately estimated.

Charu C. Agrawal et.al [21] approach work with the pseudo data rather than modification of the original data. This proposed technique is very useful for better preservation of privacy than the techniques which are uses the modification of the data.

P. Deivaini et.al [22] uses hybrid approach that is combination of the different approach which combine the given integrated results.

J. Liu et. al [23] proposed a novel algorithm which overcome the curse of dimensionality and provide the privacy to the database.

K. Alotaibi et. al[24] proposed an algorithm for multidimensional scaling that a non linear dimensionally reduction techniques to project the data into the lower dimensional.

E. Ghasemi et. al[25] proposed techniques for the trajectory data to preserve the privacy and sensitive attributes are generalized with respect to the different privacy requirement for the moving objects.

T. Jahan et. al [26] proposed a techniques for the data perturbation using the SSDV for analyzing system used to transform the original data into the distorted dataset using scarified singular value decomposition.

D. Karthikeswarant et. al[27] proposed a concept for association rule mining to sanitizes dataset using the sliding window algorithm to preserve the datasets.

M.N Kumbhar et. al[28] Association rule in horizontal and vertical distortion in which different approach of association rule is reviewed.

S. Lohiya and Lata Ragha [29] proposed a algorithm using the hybrid approach that combine the concept of K-anonymity and randomization.

Martin Beck and Michel Marc Ofer [30] proposed techniques for the anonym zing distortion for making a demonstrator user friendly interface and perform anonymization.

S. Brin et.al [31] proposed an algorithm to mine the association rules that identifies the correlation and considers both absence and presence of item sets on the basis of generating the rules that measures the significance of the association, this paper used the chi square test for correlation from the classical statics concepts.

S. Brin et. al. [32] proposed a concept of how to measures of an interest of association when the rules are generated instead of using confidence. The authors used a metric they call convocation, which measure of implication and not just co-occurrence.

5. Proposed Work

Our proposed algorithm classified into the four main levels, Level 0, Level 1, Level 2 and Level3. In Level 0 arrange all the parties in the serial manner (one after another) and assume that first party as protocol initiator party and trusted party for all the party presented in the scenario and the protocol initiator party sanded the hello massage to all the parties, as well as all the party have their own horizontally partitioned distributed database and each party have their own random number (R) for privacy purpose, when the parties received the hello massage all the parties sanded their random number to the protocol initiator party. In Level 1 assume that all the transaction and attribute values as observed frequent and convert that observed frequency to expected frequency by using the concept of data modification method (Chi Square), when all the parties calculated their chi square value they sanded to the protocol initiator party. In level 2 each party works to locally finding all frequent item sets of all the sizes by using the Apriori algorithm and all the parties calculated their own partial support by using the formula shown in algorithm2 below. Level 3 all the parties sanded their encrypted partial support value to the protocol initiator party, after that the protocol initiator party decrypted that value and calculate the global support value and broadcast that values to all the parties presented in the distributed database environments.

Algorithm2
Input: = Distributed database DB_i for i=1 to N
Process: Chi Square and Random Number
Output: Secure Global Support Value
/* Divided the entire algorithm into the four important Levels, Level 0, Level1, Level2 and Level3*/

Level0

1) Database (DB) = Set of Transactions (T);
2) Items = Position of items;
3) Transaction (T) = TID, {x | x ∈ Items};
4) Partitioned the database into the different number of parties P_i= {P_1, P_2….P_n} in horizontal partitioned approach /* $DB=DB_1+DB_2+……..+DB_n$*/.
5) All the parties are arranged in the serial manner, one after another from P_1 to P_n
6) Each party have their own random number R_i={R_1, R_2….R_n}
7) Protocol Initiator ←P_1
8) All the parties sanded their own random number to the Protocol initiator party after receiving the hello massage from the protocol initiator party
From R_2 to R_n

Level 1

9) Transaction TID (T_i){for i=1 to N }and Attribute Values A_j {for j=1 to i}=Observed Frequency (O)
10) For J=1 to N do
11) For I=1 to J do
12) Row_i Total (R_i) =$\sum_{i=1}^{j}$ Columnij
13) For I=1 to N do
14) For J=1 to I do
15) $Column_j$ Total (C_j) =$\sum_{j=1}^{i}$ Rowij
16) For I=1 to N do
17) For J=1 to N do
18) $Grand_{ij}$ Total (G_{ij})=$\sum_{i,j=1}^{n}$ DBij /*First each party calculated their sum of total number of items in the row and sum of total number of items in the particular column and after that the parties calculated sum of their grand total*/.
19) Calculate the expected frequency value by using the following formula
20) For I=1 to N do
21) For J=1 to I do

Expected Frequency (E_{ij}) = R_i (Row Total) * C_j (Column Total)/ G_{ij} (Grand Total)

22) Then calculate the value of Chi Square by using the following formula
Chi Square (κ^{2}) = (Observed Frequency (O) - Expected Frequency (E)) ^{2}/ Expected Frequency (E)
23) Observed Frequency (O)= Expected Frequency (E)

Level 2

24) Then after that each party applied the Apriori algorithm to find the list of frequent item sets of all sizes
Frequent item sets≥ Minimum Support
25) Then each party calculate their Partial support (P_i)
26) For I=1 to N do
PS_i= X_i. Support- Minimum Support*|DB|+κ_i^{2}+R_i

Level 3

27) All the parties sanded their partial support(PS) from $P_1 \leftarrow P_i$ Where I =2 to N
28) Then the protocol initiator party decrypted the encrypted partial support value
29) For I=1 to N
30) PS_i= X_i. Support- Minimum Support*|DB|-κ_i^{2}-R_i
31) Then after that the initial party calculate their global encrypt support by using the following formula
Global Support (GS) =PSi, for I=1 to N { }/N /* PSi =Partial Support, N =Total Number of Parties*/
32) Then after that protocol initiator party broadcast the global support value to all parties

In this paper, we consider the number of parties is three and each party have their own horizontally partitioned distributed database (Market Database, that contain five number of attributes) and all the database tables are given in the appendix of this paper, Table1 shows the party1 database, Table 3 shows the party 2 database and Table 5 shows the party 3 database. Party1 select their random number 1 is 5, Party2 select their random number 2 is 10 and Party3 select the random number 3 is 9 and selection of random number every parties sanded their own random number to the protocol initiator party. We consider that Table 1, Table2 and Table 3 transaction value of that particular attribute as an observed frequency, so converted that observed frequency into the expected frequency after applying the chi square concept. Table2 , Table4 and Table 6 shows the expected frequency of each parties. After that all the parties calculated their $\chi 2$ values from the expected frequency and sanded to the protocol initiator party. Then all the parties calculated the number of frequent items by applying Apriori algorithm in Table2, 4 and 6 with considering the minimum support value 0.40 for all the database tables. Calculation of expected frequency is given below for all the parties 1, 2 and 3.

Expected Frequency (E_{ij}) =Row Total* Column Total/ Grand Total

E11= 76*59/357 =12.56,E12= 77*59/357 =12.72,E13= 79*59/357=13.05,E14= 69*59/357=11.40,E15= 56*59/357=9.25,E21= 76*40/357 =8.51,E22= 77*40/357 =8.62,E23= 79*40/357=8.85,E24= 69*40/357=7.73,E25= 56*40/357=6.27,E31= 76*86/357 =18.30,E32= 77*86/357 =18.54,E33= 79*86/357=19.03,E34= 69*86/357=16.62, E35= 56*86/357=13.49, E41= 76*172/357 =36.61, E42= 77*172/357 =37.09, E43= 79*172/357=38.06, E44= 69*172/357=33.24, E45= 56*172/357=26.98

At $Party_1$ have the following frequent items set= {Beef, Chicken, Milk, Clothes, Cheese}

Chi Square (κ^{2}) = (Observed Frequency (O) - Expected Frequency (E)) ^{2}/ Expected Frequency (E)

κ^{2}=$(14-12.56)^2/12.56+(13-12.72)^2/12.72+(15-13.05)^2/13.05+(10-11.40)^2/11.40+(7-9.25)^2/9.25+(6-8.51)^2/8.51+(7-8.62)^2/8.62+(8-8.85)^2/8.85+(9-7.73)^2/7.73+(10-6.27)^2/6.27+(21-18.30)^2/18.30+(17-18.54)^2/18.54+(19-19.03)^2/19.03+(18-16.62)^2/16.62+(11-$

$13.49)^2/13.49+(35-36.61)^2/36.61+(40-37.09)^2/37.09+(37-38.06)^2/38.06+(32-33.24)^2/33.24+(28-26.98)^2/26.98$

$ϗ^2=0.16+0.006+0.29+0.171+0.54+0.74+0.304+0.081+0.20+2.21+0.39+0.127+0.0015+0.114+0.45+0.070+0.228+0.029+0.046+0.038$

$ϗ^2=6.19$

At party2 have the following frequent items set= {Beef, Chicken, Milk, Cheese}

$ϗ^2=(5-8.63)^2/8.63+(7-8.44)^2/8.44+(8-8.25)^2/8.25+(10-6.52)^2/6.52+(13-11.13)^2/11.13+(17-12.45)^2/12.45+(18-12.17)^2/12.17+(9-11.90)^2/11.90+(7-9.45)^2/9.45+(11-16.05)^2/16.05+(10-8.03)^2/8.03+(5-7.85)^2/7.85+(6-7.67)^2/7.67+(7-6.07)^2/6.07+(12-10.35)^2/10.35+(13-15.87)^2/15.87+(14-15.51)^2/15.51+(20-15.16)^2/15.16+(10-11.99)^2/11.99+(22-20.45)^2/20.45$

$ϗ^2=1.52+0.24+0.007+1.85+0.314+1.66+2.79+0.706+0.617+1.58+0.48+1.03+0.363+0.142+0.26+0.51+0.147+1.545+0.330+0.117$

$ϗ^2=16.206$

At Party3 have the following frequent items set = {Clothes, Cheese}

$ϗ^2= (6-9.76)^2/9.76+ (17-12.8)^2/12.8+(8-10.77)^2/10.77+(20-16.50)^2/16.50+(13-14.14)^2/14.14+(10-8.08)^2/8.08+(10-10.6)^2/10.6+(9-8.92)^2/8.92+(10-13.66)^2/13.66+(14-11.71)^2/11.71+(13-11.14)^2/11.14+(11-14.6)^2/14.6+(15-12.29)^2/12.29+(19-18.82)^2/18.82+(15-16.13)^2/16.13$

$ϗ^2=1.44+1.37+0.712+0.74+0.091+0.45+0.033+0.0007+0.98+0.447+0.310+0.887+0.59+0.0017+0.079$

$ϗ^2=8.13$

List of frequent item set for party 1 {Beef, Chicken, Milk, Clothes, Cheese}, List of frequent item set for party 2 {Beef, Chicken, Milk, Cheese}, List of frequent item set for party 3 {Clothes, Cheese}, In this each party calculated their partial support using the given formula adds the random number and $ϗ^2$, sanded to the protocol initiator party, partial support calculation of all the parties given below, Let the item set = {Beef, chicken} for calculation of partial support.

Partial support (PS_j) = X_j. support- Minimum Support*|DB|+$ϗ^2$+Rn

$PS_1=75.98-0.40*4+6.19+5=85.57$

$PS_2=44.98- 0.40*4+16.206+10=69.58$

$PS_3=28.98-0.40*3+8.13+9=44.91$

When all the parties calculated their partial support they sanded their encrypted partial support to the protocol initiator party, then the protocol initiator party calculated their global encrypted support by using the following formula.

Global encrypt support (GES) =Sum of all the Partial support-(Random Number + Chi Square Value)

Global encrypt support (GES) = 200.6-(6.19+16.206+8.13+5+10+9) = 200.6-54.52=146.08 ≥ 0

If the global support is greater than zero its means that the item is globally frequents item set may be its locally infrequent. And if the global support is less than zero its means that it's globally infrequent maybe it's locally frequents. Then the protocol initiator party broadcast the result to all the parties presented.

6. Conclusion and Future Work

In today world, privacy is major concern to protect their personal data, people are very much concern about their personal information which they don't want to share. In our analysis of survey papers we found that no single techniques that is consists in all domains. All the methods perform a different way depending on the types of data and type of application that they used. But still some of the method is very useful for privacy point of view, Data modification and Random number selection, so in this paper, we proposed a useful privacy preservation algorithm for mining association rules in the large business transaction data sets. The major advantage of this algorithm is that is provide the highest privacy preservation to the database because, we used the chi square to modified the personal business database and after that the each party select their own random to provide again the privacy to the database and after all the protocol initiator party broadcast the global result to all the party presented with zero percentage of data leakage. And in future we want to propose a hybrid approach that will be useful for all types of data set and any environment.

References

[1] Agrawal R., Imielinski T., Swami A.: Mining associations between sets of items in large databases, In Proc. of the ACM SIGMOD Int'l Conference on Management of Data. - Washington D.C. pp. 207-216, (1993)

[2] Agrawal R., Shrikant R.: Fast Algorithms for Mining Association rules, In Proc. Of the 20th VLDB Conference, pp. 478-499, (1994).

[3]. Fayyad U, Piatetsky-Shapiro G, Smyth P.: Knowledge Discovery and Data Mining: Towards a unifying Framework, In Proc of KDD-1996, pp 82-88, (1996).

[4]. Shintani T, Kitsurgegawa M.: Hash based Parallel Algorithm for Mining Association Rules, In Proceedings of IEEE Fourth International Conference on Parallel and Distributed Information Systems, pp.19-30, (1996).

[5]. Cohen E, Datar M, Fujiwara S, Gionis A, Indyk P, Motwani R, Ullman J and Yang C.: Finding Interesting Associations without Support Pruning, IEEE Transactions on Knowledge and Data Engineering, 13(1), pp 64 – 78, (2001).

[6]. Webb G.I.: Efficient search for association rules, International Conference on Knowledge Discovery and Data Mining, In Proc. of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts, United States, pp 99 – 107, (2000)

[7]. Hipp J. G ̈untzer U and Nakhaeizadeh G.: Algorithms for Association Rule Mining –A General Survey and Comparison, ACM SIGKDD Explorations Newsletter, 2(1), pp 58 – 64, (2000).

[8]. Palshikar G.K, Kale M.S., Apte M.M.: Association Rules Mining Using Heavy Item sets, Data & Knowledge Engineering, 61(1), pp 93-113, (2007).

[9]. Xia Y, Yang Y, Chi Y.: Mining Association Rules with Non-uniform Privacy Concerns, Data Mining And Knowledge Discovery, In Proc. of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, Paris, France, pp 27 – 34, (2004).

[10]. Wang P.: Research on privacy preserving association rule mining a survey, The 2nd IEEE International Conference on Information Management and Engineering (ICIME), Chengdu, (2010).

[11] Clifton C., Kantarcioglou M., dong Lin X. and Michaed Y.: Tools for privacy preserving distributed data mining," SIGKDD Explorations 4(2), (2002).

[12] Shang Z., Hamerlinck J.D.: Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases," Data Mining Workshops, ICDM Workshops 2007. Seventh IEEE International Conference on 28-31 pp.723–728, (2007).

[13]Chan P.: An extensible meta-learning approach for scalable and accurate inductive learning. PhD Thesis, Department of Computer Science, Columbia University, New York, NY, USA, (1996).

[14] Goldwasser S.: Multi-party computations: Past and present. In Proceedings of the 16th Annual ACM Symposium on the Principles of Distributed Computing, Santa Barbara, ACM Press, California, USA, pp.1-6, (1997).

[15]Brin S., Motwani R., Silverstein C.: Beyond market baskets: Generalizing association rules to correlations. In J. M. Peckman (ed.), Proc. ACM SIGMOD Conference on Management of Data (SIGMOD'97), pp 265-276, (1997).

[16] Liu B., Hsu W., Ma Y.: Pruning and summarizing the discovered associations. In Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD1999), pp. 125-134, (1999).

[17]Bayardo R.J., Agrawal R.: Mining the most interesting rules. In Proc. Fifth Intl. SIGKDD Conf. Knowledge Discovery and Data Mining (KDD1999), pp. 145-154, (1999).

[18]Lindell Y., Pinkas B.: Privacy preserving data mining, In proceedings of Journal of Cryptology, 5(3), (2000).

[19]. Vaidya J., Clifton C.: Privacy preserving association rule mining in vertically partitioned data, In The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining (IEEE 2002), Edmonton, Alberta, CA, (2002).

[20]Kargupta H., Datta S., Wang Q., Siva kumar K.: On the Privacy Preserving Properties of Random Data Perturbation Techniques, In proceedings of the Third IEEE International Conference on Data Mining, (2003).

[21]Aggarwal C., Yu P.S., A condensation approach to privacy preserving data mining, In proceedings of

International Conference on Extending Database Technology (EDBT), pp. 183–199, (2004).

[22]Deivanai P., Jesu Vedha Nayahi J., Kavitha V., A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data, In proceedings of International Conference on Recent Trends in Information Technology, (2011).

[23]Liu J., Luo J., Huang J.Z., Rating.: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements, In proceedings of 11th IEEE International Conference on Data Mining Workshops, (2011).

[24]Alotaibi K., Rayward-Smith V.J., Wang W., de la Iglesia B.: Non-linear Dimensionality Reduction for Privacy- Preserving Data Classification" in proceedings of 2012ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, (2012).

[25]Komishani E.G., Abadi M.: A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing, In proceedings of 6'th International Symposium on Telecommunications (IST'2012), (2012).

[26]Jahan T., Narsimha G., and Guru Rao C.V.: Data Perturbation and Features Selection in Preserving Privacy, In proceedings of IEEE (2012).

[27]Karthikeswarant D., Sudha V.M., Suresh V.M., and Sultan A.J.: A Pattern based framework for privacy preservation through Association rule Mining, In proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM -2012), IEEE 2012, (2012).

[28]Kumbhar M. N., Kharat R.: Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper, In proceedings of IEEE 2012, (2012).

[29]Lohiya S., Ragha L.: Privacy Preserving in Data Mining Using Hybrid Approach, In proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012, (2012).

[30]Beck M., Marh¨ofer M.: Privacy-Preserving Data Mining Demonstrator, In proceedings of 16th International Conference on Intelligence in Next Generation Networks, IEEE 2012, (2012).

[31]Brin, S., Motwani, R. and Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations, Proc. ACM SIGMOD Conf., pp. 265-276, (1997).

[32]Brin, S., Motwani, R., Ullman, J. D., Tsur, S., Dynamic item set counting and implication rules for market basket data. In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, pp. 255-264 (1997).

TID	Beef	Chicken	Milk	Clothes	Cheese	Total
T1	14	13	15	10	7	59
T2	6	7	8	9	10	40
T3	21	17	19	18	11	86
T4	35	40	37	32	28	172
Total	76	77	79	69	56	357

APPENDIX

Shows the horizontally partitioned distributed database table that each party have, in this paper we consider the number of parties is 3.

Table1: Party1 Have the Distributed Database

Table2: Party$_1$ have Distributed Database after conversion of observed frequency into expected frequency

TID	Beef	Chicken	Milk	Clothes	Cheese
T1	12.56	12.72	13.05	11.40	9.25
T2	8.51	8.62	8.85	7.73	6.27
T3	18.30	18.54	19.03	16.62	13.49
T4	36.61	37.09	38.06	33.24	26.98
Total	75.98	76.97	78.99	68.99	55.99

Table3: Party$_2$ have Distributed Database

TID	Beef	Chicken	Milk	Clothes	Cheese	Total
T1	5	7	8	10	13	43
T2	17	18	9	7	11	62
T3	10	5	6	7	12	40
T4	13	14	20	10	22	79
Total	45	44	43	34	58	224

Table4: Party$_2$ have Distributed Database after conversion of observed frequency into expected frequency

TID	Beef	Chicken	Milk	Clothes	Cheese
T1	E11 =8.63	E12=8.44	E13=8.25	E14=6.52	E15= 11.13
T2	E21=12.45	E22= 12.17	E23=11.90	E24= 9.41	E25=16.05
T3	E31=8.03	E32=7.85	E33=7.67	E34=6.07	E35=10.35
T4	E41=15.87	E42=15.51	E43=15.16	E44=11.99	E45=20.45
Total	44.98	43.97	42.98	33.99	57.98

Table5: Party3 have Distributed Database

TID	Beef	Chicken	Milk	Clothes	Cheese	Total
T1	6	17	8	20	13	64
T2	10	10	9	10	14	53
T3	13	11	15	19	15	73
Total	29	38	32	49	42	190

Table6: Party$_3$ have Distributed Database after conversion of observed frequency into expected frequency

TID	Beef	Chicken	Milk	Clothes	Cheese
T1	E11 =9.76	E12=12.8	E13=10.77	E14=16.50	E15= 14.14
T2	E21=8.08	E22= 10.6	E23=8.92	E24=13.66	E25=11.71
T3	E31=11.41	E32=14.6	E33=12.29	E34=18.82	E35=16.13
Total	28.98	38.00	31.98	48.98	41.98