Journal of Computer Science 5 (9): 666-673, 2009
ISSN 1549-3636

Utility Independent Privacy Preserving Data Mining on Vertically Partitioned Data

[1]E. Poovammal and [2]M. Ponnavaikko
[1]Department of Computer Science and Engineering,
Sri Ramaswamy Memorial University, Kattankulathur, Chennai-603203, India
[2]Bharathidasan University, Trichy, India

Abstract: Problem statement: Driven by mutual benefits, or by regulations that require certain data to be published, there has been a demand for the exchange and publication of data among various parties. Data publishing has been ubiquitous in many domains such as medical, business and education. Detailed person-specific data, present in the centralized server or in the distributed environment, in its original form often contains sensitive information about individuals, and publishing such data immediately violates individual privacy. The main problem in this regard is to develop method for publishing data in a more hostile environment so that the published data remains practically useful while individual privacy is preserved. There are n parties, each having a private database, want to jointly conduct a data mining operation on the union of their databases. How could these parties accomplish this without disclosing their database to the other parties or any third party? **Approach:** To address this issue, we developed a simple technique of transforming the categorical and numeric sensitive data using a mapping table and graded grouping technique, respectively. The typical data mining tasks such as classification, clustering and association rule mining were performed on both the original and transformed tables. The rules/results/patterns of both the tables were compared and the utility of the transformed data was evaluated. **Results:** The evaluation results demonstrated that the proposed approach was able to achieve cent percent utility for any type of mining task as compared to the original table. The classification accuracy of Adult data set obtained, with education as class variable was 40.08% and the same accuracy was obtained even after transformation. Similarly the number of rules generated for the given confidence 0.9, was the same for both the original and transformed table and equal to 10. **Conclusion:** The association rules involving categorical sensitive attributes were checked manually for privacy breach. We found that it is not possible to guess the actual sensitive values from the rules, even though there was no information loss. The results can be interpreted only with the concern of data owner or data publisher.

Key words: Categorical data, partitioned data, privacy preservation, sensitive data

INTRODUCTION

Progress in scientific research depends on availability and sharing of information and ideas. But protecting the privacy of human participant is given top priority by the researcher. Many privacy preserving data mining algorithms have been developed to protect the privacy of the individual even after the data mining process. Some privacy preserving data mining approaches have been developed for centralized data, while the others refer to distributed data scenario. Distributed data may be horizontally or vertically partitioned. A school is maintaining the academic records of its students in a database. Suppose a researcher wants to analyze the students' performance in an area, the academic records with the same attributes of different schools (sites) in that area are to be collected for analysis. Also, consider separate hospitals that wish to conduct a joint research while preserving the privacy of their patients.

In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party. Such kind of data with same attributes at different sites is called as horizontally partitioned data. But, if the researcher wants to find the association between the students' character and their parents occupation or between medical diagnosis and

Corresponding Author: E. Poovammal, Department of Computer Science and Engineering,
Sri Ramaswamy Memorial University, Kattankulathur, Chennai-603203, India

attendance performance, the different databases like, academic, medical, personal data of the same set of students are to be combined for analysis and such a kind of data set with a single join key (e.g., student id) is called vertically partitioned data. In other words, a portion of each instance is present at each site but no site contains complete information for any instance is vertically partitioned data.

A researcher can mine very useful rules\patterns if he is allowed to work on vertically partitioned data. For example, some cancer treatments are highly effective but have debilitating side effects with high variance between populations[1]. The factors determining the efficacy of such treatments can be learnt by decision trees\ association rules derived from vertically partitioned data tables like hospital management data, pharmacy data and insurance data, each of which is prevented by privacy laws from disclosing the individuals' identifiable information. Other than medical research, competing companies may like to perform mining tasks on data of both to get accurate results but unlike to disclose their own data to the other party. For example Ford and Firestone shared a problem with jointly produced product: Ford Explorers with Firestone tires. Factors such as trade secrets and agreements with other manufacturers stand in the way of necessary sharing. Even government entities face similar problems such as limitations on sharing between law enforcement, intelligence agencies and tax collection.

We have developed a simple technique by which vertically partitioned data can be used for any type of mining tasks. But, the individuals' privacy is preserved. Privacy preservation can mean many things: Protecting specific sensitive values of the individuals and hiding the link between the attribute values and the individuals they applied to, protecting the sources. Our goal is that by applying our technique each site can sponsor the required data to the third party, without modifying the structure of the data, so that any mining technique or algorithm, without any modification can be applied by the third party to get the actual accurate results as if mined from actual database. At the same time, the data miner can not interpret the results\patterns.

Related works: A simple approach to privacy preserving data mining over multiple sources that are not willing to share data is to apply existing techniques and tools at each site independently and combine the results. But it will not give the globally valid results because of duplicated data at different sites. Also it is not possible to detect the cross site correlations.

Another approach is to perturb the local data (by adding "noise") before the data mining process and mitigate the impact of the noise from the data mining results by using reconstruction techniques[2]. However, it is impossible to reconstruct the original data set and also the accuracy depends on the reconstruction algorithm[3]. The problem of distributed privacy preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations. Many of these techniques work by sending changed or encrypted versions of the inputs to one another in order to compute the function with different alternative versions followed by an oblivious transfer protocol to retrieve the correct value of the final output. The algorithms for secure multiparty computation over horizontally partitioned data set include Naïve Bayes classifier[4], Support Vector Machine (SVM) classifier with non linear kernels[5], Association Rule Mining[6], Clustering[7-9].

The approach of vertically partitioned mining has been extended to a variety of data mining application such as Naïve Bayes classifier[10], SVM classification[11] ,decision trees[12] K-means clustering[13] and Association Rule Mining[14,15]. Vaidya and Clifton[16] gave a nice algebraic solution for vertically partitioned data. However, this solution can leak many linear combinations of each party's private data to other. Also, to process one candidate frequent item set, its computational overhead is quadratic in the number of transactions. Two algorithms are given by Sheng Zhong in[17] which are having computational overheads linear to the number of transactions. But when his technique is used in practice, it should be complemented by other algorithm that computes all frequent item sets without testing candidates one by one.

All of the cryptographic work falls under the theoretical framework of Secure Multiparty Computation. In Agrawal's study[2] the privacy-preserving data mining problem between two parties is solved by data perturbation method while Lindell and Pinkas use secure multi-party computation protocols[18] to solve the problem. We have proposed a framework that allows us to systematically transform normal data mining computations to secure multi-party computations. The problem is defined as this: There are n parties, each having a private database, want to jointly conduct a data mining operation on the union of their databases. How could these parties accomplish this without disclosing their database to the other parties or any third party?

MATERIALS AND METHOD

The framework of our privacy preserving mining model is as shown in Fig. 1.

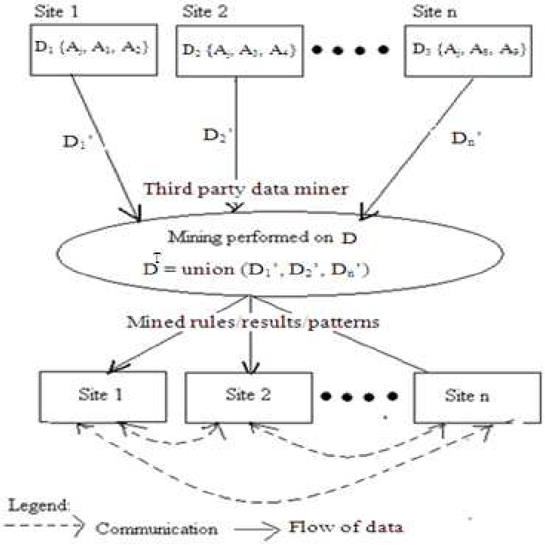


Fig. 1: Frame work of vertically partitioned privacy preserving model

Our model assumes that all sites collect data for the exact same set of entities. The assumption can be neglected by deciding the behavior for missing values. For example, missing value may be replaced by the average if it is numerical data or by mostly used value if it is categorical data. Based on this assumption, attribute A_j is common to all the vertically partitioned data sets (D_1-D_n) and hence form the join key. Also the number of rows is almost same in all the sites.

Data flow: All the n parties available in Site 1 to Site n have their own datasets D_1-D_n with only one attribute A_j in common, called join key attribute. In some situations only a part of the data set needs to be kept confidential. These attributes are sensitive attributes and all the other attributes don't need any treatment.

All the parties want to jointly conduct data mining operation on a single database D which is formed by the union of all the datasets $D_1\{A_j, A_1, A_2,\ldots\}$, $D_2\{A_j,A_3, A_4,\ldots\}$… and $D_n\{A_j,A_8, A_9,\ldots\}$ to get better results. But to preserve the privacy of the actual values of the individual databases, the third party data miner is allowed to work with a single database D^T which is formed by the union of all the transformed data bases D_1', D_2'…D_n' where $D_1' = \{A_j, A_{1T}, A_{2T}\ldots\}$ $D_2' = \{A_j, A_{3T}, A_{4T}\ldots\}$….and $D_n' = \{A_j, A_{8T},A_{9T}\ldots\}$ where A_{xT} is the transformed value of the sensitive attribute A_x.

An attribute is called Sensitive, if the individual is not willing to disclose or an adversary must not be allowed to discover the value of that attribute. The method of converting the attribute A_x to A_{xT} is explained in the next section.

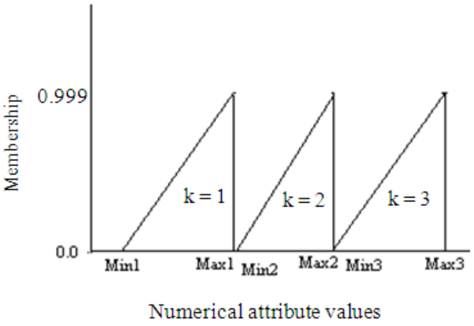


Fig. 2: Graded grouping

The data miner who works on the transformed single data base D^T can perform any data mining task, as if he works on the original data. But he can not interpret the results\rules\patterns. He can declare the results to all the parties who have participated in the data sharing. The individual parties can interpret a few results, which contain the transformed attribute values of their own data bases. To interpret the remaining results, each site should communicate with the other sites and mutually exchange the actual values, involved in the results. Since, the actual value of the results\rules\patterns alone known by all the parties, the individuals' privacy is preserved.

Transformation: The transformation of the attribute A_x-A_{xT} is based on the data type of attribute A_x. If the data type is numerical we use graded grouping technique and if the data type is categorical we use mapping table for the transformation.

Graded grouping technique: This is a simple transformation method which maintains the correlation factor of nearly 1 between the transformed values and the original values. Our approach to numerical attribute is graded grouping is as shown in Fig. 2. To convert the actual values of a single numeric attribute, the following steps are followed. First step is to fix the number of categories (k) for the given range. Second step is, for each category C_1 …C_k, the max and min value is to be fixed in such a way that non overlapping continuous range results. Range for each category may or may not be uniform. If the uniform range is considered for each category then the correlation factor between the original and transformed values is one. Otherwise, it will decrease but maintains a positive linearity. Third step is to fix the category (Ci) for each actual value(x), to which it belongs and find the membership value m(x) using:

$m(x) = 0.0$ if $x = \min(Ci)$
$= (x–\min(C_i))/(\max(C_i)-\min(C_i))$
if $\min(Ci)>x<\max(Ci)$
$= 0.999$ if $x = \max(Ci)$

The fourth step is to replace the actual value with a new value n(x) or transformed value, which can be calculated by adding category number (Ci) and the membership value m(x).

Algorithm for graded grouping: Function graded_grouping (A^n).

Input: n records of numerical data type (actual values)
Output: n records of numerical data type (transformed values)

```
1.  Get the value of k \\ Number of categories fixed by
    the individual site
2. For i = 1 to k
    Get min(Ci) and max(Ci) \\ Range for each
    category fixed by the individual site
3. For j = 1 to n \\ number of records = n
    Let i = 1
     Do while i< =k
        If min(Ci) ≤ A[j] ≤ max(Ci)
              CX[j] = i
        Else i++
        Endif
    End while
    If A[j] = min(Ci)
    MX[j] = 0.0
        Else If A[j] = max(Ci)
        MX[j] = 0.999
        Else
        MX[j] = (A[j]-min(Ci))/(max(Ci)-min(Ci))
              NX[j] = MX[j]+CX[j]
        End if
    End for
```

Mapping table: For the categorical attribute, all the values are given an alias names and the original values are mapped to alias names in a mapping table. The mapping table is preserved by the individual sites. So, the transformed categorical values contain only the alias names and shared with the third party data miner, who can not interpret any actual values. An example mapping table is shown in Table 1.

Experimental setup: The data miner's job is to perform union operation on the various transformed attributes A_{1T}, A_{2T}… A_{9T} and other non sensitive attributes using the join key attribute A_j to form a single table D^T, which alone can be used for any data mining task. We have decided to conduct the experiment on real data set and hence used the adult database from UCI machine learning repository[19] with 35,561 records.

Table 1: Mapping table for categorical sensitive attribute education

Actual value	Transformed Value
Bachelors	Education_1
HS-grad	Education_2
11th	Education_3
Masters	Education_4
9th	Education_5
Some-college	Education_6
Assoc_acdm	Education_7
7th-8th	Education_8
Doctorate	Education_9
Assoc-VOC	Education_10
Prof-school	Education_11
5th-6th	Education_12
10th	Education_13
1st-4th	Education_14
Preschool	Education_15
12th	Education_16

The attributes Age, Work class, Education, Marital status, Occupation, Relation, Race, Sex, country (form table D) are considered for analysis, assuming that different attributes are received from different sites. We considered age as sensitive numerical attribute and hence Age_T calculated by our algorithm for graded grouping. Similarly, education is considered as sensitive categorical attribute and hence $Education_T$ is formed from the mapping table shown in Table 1. We have implemented the algorithm in Java standard Edition 5.0 and made to run on Intel® Core2 Duo, 1.8 GHz, 1 GB RAM system which took only 28sec for generating privacy preserving Adult data set D^T. The various data mining tasks on both table D and new table D^T with the attributes Age_T, Work class, $Education_T$, Marital status, Occupation, Relation, Race, Sex, country are performed using the tool WEKA[20] and the results are compared.

RESULTS

For evaluation purposes, we performed the mining tasks such as classification, association rule mining and clustering on both the original adult data (D) and transformed table (D^T). The results were compared. Classification was performed by decision tree (J48) method and zeroR method, considering education as classification variable. Parameters compared are shown in Table 2. The results were not affected by the proposed transformation method. In J48 method highest F-measure value was 0.874, for doctorate in the original table but for Education_9 in the transformed table.

Table 2: Comparison of results for different classification schemes

	Classification scheme	
Parameters	Weka.classifiers.trees.J48	Weka.classifiers rules.ZeroR
Test mode	10-fold cross-validation	10-fold cross-validation
Correctly classified instances	Original/transformed table -13051 (40.0854 %)	Original/transformed table-0499 (32.2471%)
Incorrectly classified instances	Original/transformed table -19507 (59.9146 %)	Original/transformed table-2059 (67.7529%)
Kappa statistic	Original/transformed table -0.1924	Original/transformed table-0
Mean absolute error	Original/transformed table -0.0932	Original/transformed table-.1012
Root mean squared error	Original/transformed table -0.2183	Original/transformed table-0.2249
Relative absolute error	Original/transformed table-92.063 %	Original/transformed table-100%
Root relative squared error	Original/transformed table-97.063 %	Original/transformed table-100%
Time taken to build model	Original table-0.61 sec Transformed table-1.06 sec	Original/transformed table-0.11 sec
Highest F-measure for class	Original table-0.874 (for doctorate) Transformed table-0.874 (for Education_9)	Original table-0.488 (for HS_Grad) Transformed table-0.488 for Education_2

Table 3: Comparison of results for different association rule mining schemes

	Association rule scheme	
Parameter	Weka.associations apriori	Weka.associations tertius
No. of rules	Original/transformed table-10 (confidence> = 0.91)	Original/transformed table-11(confidence> = 0.95)

Table 4: Comparison of results for different clustering schemes

	Clustering scheme	
Parameters	Density based clusterer	Simple K means euclidean distance
Within cluster sum of squared errors	Original/transformed table-92466	Original/transformed table-92466
Instances cluster 0	Original/transformed table-8480 (57%)	Original/transformed table-20672 (63%)
Instances cluster 1	Original/transformed table-4078 (43%)	Original/Transformed table-1886 (37%)

So, any data miner working on the transformed table can not guess the actual value, without mapping table. Table 3 and 4 show the comparison results of association rule mining and clustering.

DISCUSSION

Any data mining technique can be evaluated by the parameters like performance of algorithm, data utility after transformation, level of uncertainty, resistance accomplished by the privacy algorithms[21].

Performance of algorithm: The performance of any privacy algorithm can be measured by the time needed to hide a specified set of sensitive information. We have considered the original Adult Data set table, with two sensitive attributes age (Numeric) and Education (Categorical), as the table for transformation. Our algorithm took only 28sec for transformation. Time complexity of our algorithm is linear to size of the table.

Data utility after transformation: The data utility after transformation can be measured by the parameter, loss of functionality or information loss. For example, suppression and generalization are some form of transformation. If suppression is used for an attribute value, utility of that data gets reduced, since missing values can not be handled by many mining tools.

Use of sampling does not modify the information stored in the data base, but still utility gets reduced, since information is not complete. The measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed. For example in the case of association rule mining information loss can be measured by counting the number of rules framed for the given support and confidence, before and after transformation. From the Table 3, we conclude that information loss is nil because, from the original and transformed table we get the same number of rules, whatever may the type of algorithm used.

Level of uncertainty: The level of uncertainty is the measure of capability of predicting hidden data from the data se, given for analysis or from the rules/results declared. For example, randomization is the method used to hide the data. To maintain the information, if the randomization is done to have correlation very close to one, then the data reconstruction procedure, discloses the actual values, otherwise there is loss of information. But in our method, numerical attribute preserves the information while the actual values can not be guessed without knowing the number of categories and the range of each category. Similarly, without access to the mapping table actual value can not be guessed. Consider the snap shots of Association Rule Mining experiment conducted on D and D^T by Tertius method shown in Fig. 3 and 4. The number rules are the same in both the cases.

Also the rules are exactly the same except for Rule number 6 which contains the sensitive attribute Education (Fig. 3) with the value preschool. But the same rule number in transformed table (Fig. 4) contains the sensitive attribute H_Education with the value Education_15 which no one can interpret except the owner of data set, with the availability of mapping table. Since, WEKA Association Rule mining tool can not handle numeric attribute, attribute Age is not considered for analysis.

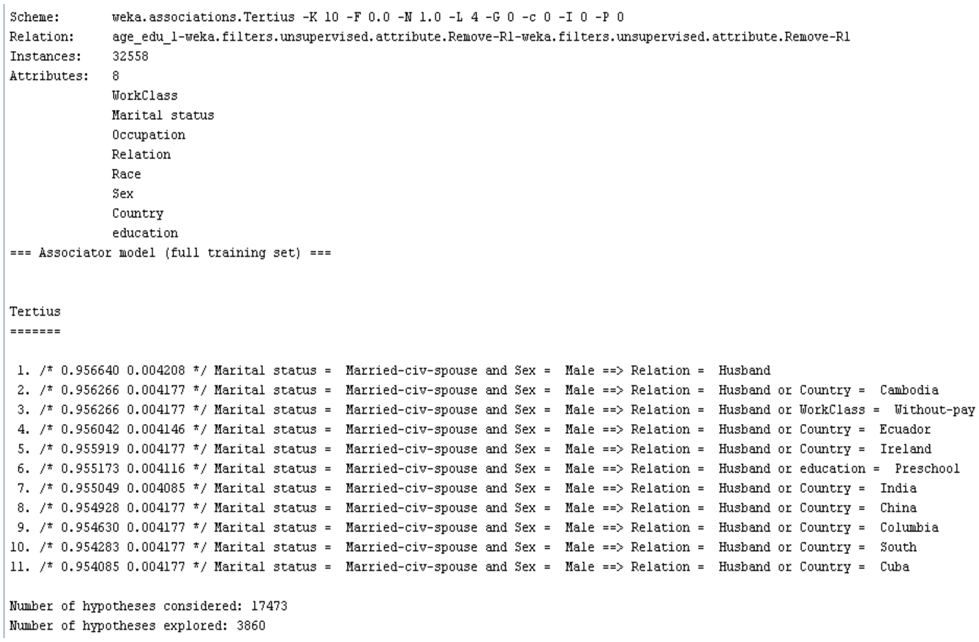
Resistance accomplished: If the resistance accomplished by the privacy algorithm is low, means that the sanitization algorithm developed against a particular data mining technique that assures privacy of information may not attain similar protection against all possible data mining algorithm. But, the resistance of our algorithm is high enough so that, whatever may be the mining task performed, the sensitive information does not leak out. For example association rule mining using Predictive Apriori algorithm was performed on the original and the transformed table. The time taken by the task for the original and transformed table are 53 min 45 sec and 54 min 38 sec respectively, while the number of rules framed in both the cases is the same for the given confidence. The fourth rule framed by this task is shown in Table 5.

Limitation: We assume that the parties participating in the process are honest but only curious to know about others.

Once the rules are declared by the data miner, being honest, the parties should be giving the actual values corresponding to the transformed values, (if they have) to other parties. Since each party knows its own data and resultant association rules, there may be some information disclosure. For example, the support of the Rule A -> B is 10% and it is known by both the parties. If Item set A and item Set B belongs to Party I and Party II respectively, who participated in the mining task, then they can calculate the value of the opponent's item set. But it is acceptable to disclose knowledge that could be obtained from global rules.

Table 5: Predictive apriori method: A sample rule

Original table	Transformed Table:
Rule 4) workClass = Private, Occupation = Sales, relation = Husband, Education= Bachelors (298) ==> Marital status = Married -civ-spouse, Sex = Male (298) acc:(0.99499)	Rule 4) workClass = private, Occupation = Sales, relation = Husband, H_Education = Education _1 (298) ==> Marital status = Married-civ-spouse, Sex = Male (298) acc:(0.99499)



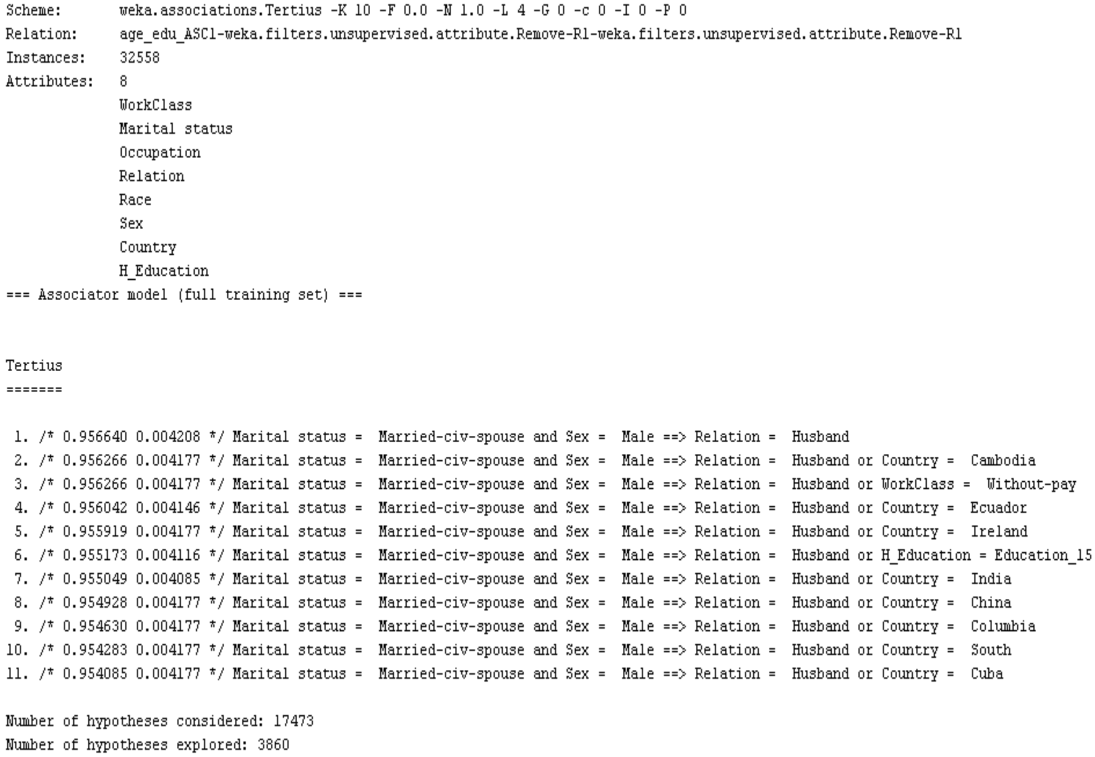
```
Scheme:       weka.associations.Tertius -K 10 -F 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -P 0
Relation:     age_edu_1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    32558
Attributes:   8
              WorkClass
              Marital status
              Occupation
              Relation
              Race
              Sex
              Country
              education
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.956640 0.004208 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband
 2. /* 0.956266 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Cambodia
 3. /* 0.956266 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or WorkClass =  Without-pay
 4. /* 0.956042 0.004146 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Ecuador
 5. /* 0.955919 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Ireland
 6. /* 0.955173 0.004116 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or education =  Preschool
 7. /* 0.955049 0.004085 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  India
 8. /* 0.954928 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  China
 9. /* 0.954630 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Columbia
10. /* 0.954283 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  South
11. /* 0.954085 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Cuba

Number of hypotheses considered: 17473
Number of hypotheses explored: 3860
```

Fig. 3: Rules generated by Tertius association rule mining technique on original table using WEKA



```
Scheme:       weka.associations.Tertius -K 10 -F 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -P 0
Relation:     age_edu_ASC1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    32558
Attributes:   8
              WorkClass
              Marital status
              Occupation
              Relation
              Race
              Sex
              Country
              H_Education
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.956640 0.004208 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband
 2. /* 0.956266 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Cambodia
 3. /* 0.956266 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or WorkClass =  Without-pay
 4. /* 0.956042 0.004146 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Ecuador
 5. /* 0.955919 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Ireland
 6. /* 0.955173 0.004116 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or H_Education = Education_15
 7. /* 0.955049 0.004085 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  India
 8. /* 0.954928 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  China
 9. /* 0.954630 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Columbia
10. /* 0.954283 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  South
11. /* 0.954085 0.004177 */ Marital status =  Married-civ-spouse and Sex =  Male ==> Relation =  Husband or Country =  Cuba

Number of hypotheses considered: 17473
Number of hypotheses explored: 3860
```

Fig. 4: Rules generated by Tertius association rule mining technique on transformed table using WEKA

CONCLUSION

Many works limited to Boolean association rule mining. But, Non categorical attributes and quantitative association rule mining are significantly more complex but using our algorithm they can be handled easily. Our goal is to develop methods enabling any data mining tasks that can be done at a single site to be done across various sources, while respecting their privacy policies and is achieved. Transformation can be easily implemented at the data source itself, whatever may be the number of sensitive attributes, at the user machine. This increases the confidence of the user in providing accurate information since he/she does not have to trust a third party to carry out the transformation process. Also many techniques concern about output privacy, whereas our focus is on the privacy of input data given for mining. Mining the distributed database can be significantly more expensive in terms of both time and space as compared to mining the true data base[22]. We have treated the distributed data as centralized data, before any mining tool is applied and hence time taken for mining is reduced.

REFERENCES

1. Shirao, K., P. Hoff, A. Ohtsu, P. Loehrer and F. Abbruzzese *et al*., 2004.Comparison of the efficacy, toxicity and pharmacokinetics of a uracil/tegafur (UFT) plus oral Leucovorin (LV) regimen between Japanese and American patients with advanced colorectal cancer: Joint United States and Japan study of UFT/LV. J. Clin. Oncol., 22: 3466-3474. DOI: 10.1200/JCO.2004.05.017
2. Agrawal, R. and R. Srikant, 2000. Privacy-preserving data mining. ACM. SIGMOD. Rec., 29: 439-450. http://portal.acm.org/citation.cfm?id=335438
3. Agrawal, D. and C.C. Aggarwal, 2002. On the Design and quantification of privacy-preserving data mining algorithms. Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, (PDS'02), ACM Press, Santa Barbara, California, US., pp: 247-255. http://portal.acm.org/citation.cfm?id=375602
4. Kantarcioglu, M. and J. Vaidya, 2003. Privacy preserving naïve bayes classifier for horizontally partitioned data. Proceedings of the IEEE 2nd Workshop on Privacy Preserving Data Mining, Nov. 19, Melbourne, Florida, USA., pp: 1-7. http://www.cis.syr.edu/~wedu/ppdm2003/papers/1.pdf

5. Yu, H., X. Jiang and J. Vaidya, 2006. Privacy Preserving SVM using nonlinear Kernels on horizontally partitioned data. Proceedings of the 2006 ACM Symposium on Applied Computing Apr. 23-27, ACM Press, Dijon, France, pp: 603-610. http://portal.acm.org/citation.cfm?id=1141277.1141415
6. Kantarcioglu, M. and C. Clifton, 2004. Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE. Trans. Knowl. Data Eng. J., 16: 1026-1037. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01316832
7. Inan, A., Y. Saygin, E. Savas, A. HIntoglu and A. Levi, 2006. Privacy preserving clustering on horizontally partitioned data. Proceedings of the 22nd International Data Engineering Workshops, (IDEW'06), IEEE Xplore Press, Atlanta, GA., USA., pp: 95-99. DOI: 10.1109/ICDEW.2006.115
8. Jagannathan, G. and R. Wright, 2005. Privacy preserving distributed K-means clustering over arbitrarily partitioned data. Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining**,** Aug. 21-24, ACM Press, Chicago, Illinois, USA., pp: 593-599. http://portal.acm.org/citation.cfm?id=1081942
9. Jagannathan, G., K. Pillaipakkamnatt and R. Wright, 2006. A new privacy preserving distributed k-clustering algorithm. Proceedings of the SIAM Conference on Data Mining, April 20-22, Bethesda, Maryland, pp: 494-498. http://www.siam.org/proceedings/datamining/2006/dm06_048jagannag.pdf
10. Vaidya, J. and C. Clifton, 2004. Privacy preserving naïve bayes classifier over vertically partitioned data. Proceedings of the SIAM International Conference on Data Mining, (DM'04), Lake Buena Vista, Florida, pp: 522-526. http://www.siam.org/meetings/sdm04/proceedings/sdm04_059.pdf
11. Yu, H., J. Vaidya and X. Jiang, 2006. Privacy preserving SVM classification on vertically partitioned data. Proceedings of the PAKDD Conference, Apr. 9-12, Singapore, pp: 647-656. http://cat.inist.fr/?aModele=afficheN&cpsidt=19687511
12. Vaidya, J. and C. Clifton, 2005. Privacy preserving decision trees over vertically partitioned data. Lecture Notes Comput. Sci., 3654: 139-152. http://cat.inist.fr/?aModele=afficheN&cpsidt=17026943
13. Vaidya, J. and C. Clifton, 2003. Privacy preserving K-means clustering over vertically partitioned data. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 24-27, ACM Press, Washington DC., pp: 206-215. http://portal.acm.org/citation.cfm?doid=956750.956776
14. Ioannidis, I., A. Grama and M. Atallah, 2002. A Secure protocol for computing dot products in clustered and distributed environments. Proceedings of the International Conference on Parallel Processing, (PP'02), IEEE Xplore Press, USA., pp: 379-384. DOI: 10.1109/ICPP.2002.1040894
15. Clifton, C., M. Kantarcioglou, X. Lin and M. Zhu, 2002. Tools for privacy preserving distributed data mining. ACM. SIGKDD. Explorat., 4: 28-34. http://portal.acm.org/citation.cfm?id=772862.772867
16. Vaidya, J. and C. Clifton, 2002. Privacy preserving association rule mining in vertically partitioned data. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, ACM Press, Edmonton, Alberta, Canada, pp: 639-644. http://portal.acm.org/citation.cfm?id=775047.775142
17. Sheng Zhong, 2007. Privacy preserving algorithms for distributed mining of frequent item sets. J. Inform. Sci., 177: 490-503. DOI: 10.1016/j.ins.2006.08.010
18. Lindell, Y. and B. Pinkas, 2002. Privacy preserving data mining. J. Cryptol., 15: 177-206. DOI: 10.1007/s00145-001-0019-2
19. Newman, D.J., S. Hettich, C.L. Blake and C.J. Merz, 1998. UCI repository of machine learning databases. University of California, Irvine. http://archive.ics.uci.edu/ml/
20. Ian, H. Witten and Eibe Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, ISBN: 0120884070, pp: 525.
21. Verykios, V.S., E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, 2004. State-of-the-art in privacy preserving data mining. SIGMOD Rec., 33: 50-57. http://portal.acm.org/citation.cfm?id=974131
22. Rizvi, S.J. and J.R. Haritsa, 2002. Maintaining data privacy in association rule mining. Proceedings of the 28th International Conference on Very Large Data Bases, Aug. 20-23, ACM Press, Hong Kong, China, pp: 682-693. http://portal.acm.org/citation.cfm?id=1287428