ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Rough set based Privacy Preserving Attribute Reduction on Horizontally partitioned data and generation of Rules

N.V.S.Lakshmipathi Raju[1], Dr. M.N. Seetaramanath[2], Dr. P. Srinivasa Rao[3], G.Nandini[4]

Assistant Professor, Department of CSE, Gayatri Vidya Parishad College of Engineering, Visakhapatnam, India[1]

Professor, Department of IT, Gayatri Vidya Parishad College of Engineering, Visakhapatnam, India[2]

Professor, Department of CS&SE, AU College of Engineering, Visakhapatnam, India[3]

Student, Department of CSE, Gayatri Vidya Parishad College of Engineering, Visakhapatnam, India[4]

Abstract: Data mining is the process of extracting knowledge from various databases. For security reasons organizations may partition their data horizontally or vertically. Another situation is hospitals maintaining sensitive information about patients. This can also be treated as horizontal partitioning of data if all the hospitals maintain data in the same format. Secure multiparty technique is a cryptographic method which can be used to derive useful results from partitioned data bases without violating the privacy of individuals or organizations. This paper describes a method for finding a rough set concept called reduct set from the partitioned databases using secure multiparty technique. From this reduct set, rules are generated using Naïve Bayesian algorithm. This paper also describes attribute reduction technique for horizontally partitioning databases .Generally it is necessary to implement data preprocessing operations on the data bases. This paper also describes an enhanced Transitive Closure algorithm for data cleaning.

Keywords: Rough sets;attribute reduction;data cleaning and rule generation

I. INTRODUCTION

Data mining is one area of science, which is used to deal with huge amounts of data. Besides from being huge in quantities, data is also distributed between many parties thereby making it even more difficult to manage. Data collaboration is an important task to generate the new rules, these rules are very useful to researchers for their research activities. The data collaboration can be done by without revealing the privacy of the involved parties. Hence privacy preservation becomes a very important task to be taken care of as no party wants to disclose its private data to anybody else[1].

The concept of Secure Multiparty Computation (SMC) is used to solve the problem of Privacy Preserving Data Mining (PPDM).This paper consider horizontal partitioning databases. Secure Multiparty Computation algorithm is used to combine the horizontal partitioning databases[1]. This algorithm is used to combine the input data bases without revealing their original datasets, which generates a reduct set. Reduct set is a concept of rough set theory[5], which was introduced by Z.Pawlak. Rough set theory is a powerful tool in the field of data mining. The notion of reduct plays a key role in rough set-based attribute reduction. In rough set theory, a reduct is generally defined as a minimal subset of attributes that can classify the same domain of objects as unambiguously as the original set of attributes.

This paper describes an algorithm for Data cleaning and rule detection. Naïve Bayesian classification approach is used to generate rules. Introduce some concepts of rough sets and Secure Multiparty encryption technique that is used in this paper.

A.Rough sets

Information System and Decision table:

In rough set, an information system is a representation of data that describes some objects. An information system S is composed of a 4-tuple $S = <U, A, V, f >$, where U is a closed universe of N objects $\{x1, x2,..., xN\}$, a non-empty

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

finite set; A is a non-empty finite set of n attribute $\{a1,a2,...,an\}$;$V=\cup Va$ where Va is the value of the attribute a; $f: U \times A \rightarrow V$ is the total decision function called the information function such that $f(x,a) \in Va$ for every $a \in A$, $x \in U$. If the attribute set A is divided into two disjoint sets, conditional attribute set C and decision attribute set D, such that $C \cup D = A$ and $C \cap D = \emptyset$, then the special information system is called a decision table or decision system [5].

Knowledge Granularity Definition:

For decision table $S=(U,C \cup D,V,f)$, if $B \subseteq (C \cup D)$ and $U/B=\{X1,X2,\ldots,Xm\}$. Granularity of knowledge of B is

$$GK(B)=\sum_{i=1}^{m}\frac{|X_i|^2}{|U|^2} \quad \ldots\ldots(1)$$

defined by[1][2]

Relative Granularity Definition:

For decision table $S=(U,C \cup D,V,f)$, if P $(C \cup D)$ and Q $(C \cup D)$, *relative granularity* of P with reference to Q is defined by[1][2]

$$GK(Q|P)=GK(P)-GK(P \cup Q) \quad \ldots\ldots (2)$$

There has been work in distributed attribute reduction that does not consider privacy issues [1].

Core Definition:

For decision table $S=(U,C \cup D,V,f)$, we have the core attributes Core$C(D)$ as followings:

$$\text{Core}C(D)=\{b \varepsilon C|GK(D|C-\{b\})-GK(D|C)>0\} \ldots..(3)$$

Significance Definition:

For decision table S=(U,C∪D,V,f), R⊆C, a∈C-R, the significance of attribute a is defined as:

$$SIG(a,R,D)=GK(D|R)-GK(D|R \cup \{a\}) \ldots\ldots.(4)$$

B.Secure Multiparty Computation:

The concept of Secure Multi-party Computation (SMC) encryption technique is used for Horizontal partitioning data bases[1].The main objective of SMC is that a computation is more secure, no party knows anything except its own input and the results. This model maintains a trusted third party, every participants gives their input to the trusted third party. Trusted third party performs computations on the inputs and sends the results to participants. For example Alice and bob sends their input data to the trusted third party. Trusted third party calculates the reduct sets from the given two databases and finally it sends the results to both the participants without bleaching the privacy of one to another.

This paper is organized as follows: Section II. Existing System. Section III Proposed system. Section IV Results. Conclusion and Future Work of this paper is presented in Section V.

II.Existing System:

Attribute Reduction Algorithm:

Let us consider the schemes based on relative granularity. Let *Sa* represent Alice's data set, and let *Sb* represent Bob's data set. The following is the attribute reduction algorithm on (*Sa*, *Sb*) as following:

Input: Alice's decision table $Sa=(Ua,C \cup D,V,f)$, Bob's decision table $Sb=(Ub,C \cup D,V,f)$.
Output: a relative reduction of $Sa \cup Sb$.
1. let $U=Ua \cup Ub$, $R=\emptyset$, calculate relative granularity$GK(D|C)$
2. for each attribute $b \in C$ do
3. Calculate relative granularity $GK(D|C-\{b\}))$
4. If $GK(D|C-\{b\})>GK(D|C)$, then $R=R \cup \{b\}$
5. End for
6. Core$C(D)=R$, if $R \neq \emptyset$then
7. Calculate relative granularity $GK(D|R)$
8. If $GK(D|R)=GK(D|$C), then go to 17
9. End if
10. Let Core$C(D)=R$
11. For each attribute $a \in C-R$, calculate $GK(D|R \cup \{a\})$
12. Choose the attribute making the value of $GK(D|R \cup \{a\})$ Minimal, which is denoted by $a0$ (If there are many, choose the one making the value of $|U/R \cup \{a\}|$ minimal as $a0$), and $R=R \cup \{a0\}$
13. Calculate the relative granularity $GK(D|R)$, if $GK(D|R)\_GK(D|C)$, then go to (11)
14. For each attribute $a \varepsilon R$-Core$C(D)$ do
15. If $GK(D|R-\{a\})=GK(D|C)$, then $R=R-\{a\}$
16. End for
17. Output R is a relative reduction of C

Here Knowledge Granularity [1] is computed for calculating the reduct sets.

III. Proposed System:

Data cleaning and Rule Detection:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Let us consider the real world phenomenon, where the data is distributed between two parties, they are Alice and Bob represented by Ua and Ub respectively**.**

Let S_a represent Alice's dataset, and let S_b represent Bob's data set. This paper uses an attribute reduction algorithm to compute the reduct set for horizontal partitioning databases [1]. There are some protocols with Secure Multi-party computation encryption technique. This paper implements a securely computing scalar product protocol for horizontal partitioning databases. This is used to generate the reduct sets from the given databases. This paper generates new rules from the obtained reduct set by using a Naive Bayesian classification algorithm. Databases should be consistent, and then only it is possible to obtain the better reduct set from it. For this, it is necessary to implement a data preprocessing operations on the databases[6]. This paper generates an enhanced transitive closure algorithm to remove duplicate data, inconsistent data and erroneous data from the user databases.

Proposed Algorithm:

```
//To extract primary keys
        For a  every query in  rs
  Loop begins;
       If  string 1  consists of null values
            then  print  unique  key  inconsistency  and  delete
that particular inconsistent record.
  else
       Then   print primary key of string 1
Loop Ends;
      If string 2 consists of unique values
               put string and primary key in Hash table
//Finding the Missing Values
       for every column count
              if rs4 = stmt5.executeQuery where column name
IS NULL OR not;
             if (rs4.getInt(1)==0)
        print no missing values

    else

       Tuples missing and column names are printed

// to Clean the records

     if(rs2.getInt(1)>1)

        Select primary key for he given string

     if  records  are  repeated  then  delete  records  due  to
redundancy

   if record is not present

       then put  the record is not found in hash table.

Delete missing values

If (rs3.getInt(colum)≥50%)

    then Delete that  record

If (rs3.getInt(colum)≤50%)
```

then replace all the missing values with either its attribute mean or global constant in the case of numerical attributes and replace with its most probable data in the case of categorical attribute.

This algorithm considers the following conditions to delete duplicate data, inconsistent data and noisy data. The following conditions identifying the given data as a duplicate data. This algorithm removes such type of tuples from the database.

[a] 2 matches if at least one of them is a primary key.

[b] 3 matches if at least two are secondary keys.

[c] 4 matches if at least one key is a secondary key.

Database after performing Data Cleaning on Horizontal partitioning databases:

Table1: Alice

DAY	OUTLOOK	TEMP	HUMIDITY	WINDY	PLAYBALL
d1	Sunny	hot	High	weak	No
d2	Sunny	hot	High	strong	No
d3	Sunny	hot	High	strong	No
d4	Sunny	warm	Strong	weak	Yes
d5	Rain	warm	High	weak	Yes
d6	Rain	cool	Normal	weak	Yes
d7	Sunny	warm	Strong	weak	Yes
d8	Sunny	warm	High	weak	Yes
d9	Rain	cool	High	strong	No
d10	Cloudy	high	Cool	strong	No
d11	Cloudy	cool	normal	strong	Yes
d12	Sunny	high	hot	strong	Yes
d13	Cloudy	warm	high	weak	Yes
d14	Rain	cool	high	normal	No
d15	Cloudy	warm	high	weak	Yes
d16	Sunny	warm	high	weak	Yes
d17	Rain	cool	high	Strong	Yes
d18	Sunny	warm	strong	weak	Yes
d19	Rain	warm	high	strong	No
d20	Rain	hot	high	weak	No
d21	Sunny	cool	high	strong	Yes
d22	Sunny	warm	strong	weak	Yes
d23	Cloudy	hot	high	weak	Yes

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Table2: BOB

DAY	OUTLOOK	TEMP	HUMIDITY	WINDY	PLAYBALL
d1	Cloudy	cool	normal	strong	Yes
d2	Sunny	warm	high	weak	No
d3	Sunny	cool	normal	weak	Yes
d4	Sunny	warm	normal	weak	Yes
d5	Sunny	warm	normal	strong	yes
d6	Cloudy	warm	high	strong	yes
d7	Cloudy	hot	normal	weak	yes
d8	Cloudy	hot	normal	weak	yes
d9	Rain	warm	high	strong	no
d10	Rain	cool	normal	strong	no
d11	Sunny	hot	high	weak	no
d12	Cloudy	hot	high	strong	no
d13	Sunny	warm	normal	weak	yes
d14	Rain	cool	normal	weak	yes
d15	Cloudy	cool	normal	weak	yes
d16	Rain	warm	high	weak	yes
d17	Rain	cool	normal	strong	no
d18	Sunny	warm	normal	weak	yes
d19	Rain	warm	normal	strong	yes
d20	Rain	warm	high	strong	no
d21	Sunny	warm	high	weak	yes
d22	Sunny	warm	normal	weak	yes
d23	Sunny	warm	normal	weak	yes

By using the following equations, the knowledge granularity can be actually computed using the scalar product protocol in Horizontal Partitioned data [1].

$$\sum_{i=1}^{n}\frac{|X_i \cup Y_i|^2}{|U|^2}=\sum_{i=1}^{n}\frac{(|X_i|+|Y_i|)^2}{|U|^2}$$
$$=\sum_{i=1}^{n}\frac{|X_i|^2}{|U|^2}+\sum_{i=1}^{n}\frac{|Y_i|^2}{|U|^2}+\frac{2(|X_1|,\cdots,|X_n|)\bullet(|Y_1|,\cdots,|Y_n|)}{|U|^2} \quad (5)$$

IV. RESULTS:

This paper uses an enhanced transitive closure algorithm for data cleaning on original data base. This technique is used to remove the noisy data, redundant data and inconsistent data from the original database. Attribute Reduction algorithm is used to find out the reduct set in Horizontal Partitioning databases.

The following are the reduct sets obtained from the Horizontal partitioning databases without using data cleaning algorithm.

```
{Temparture,Humidity,Windy},
{Humidity,Windy},
{Windy}{Humidity},
{Temperature,Windy},
{Temperature},
{Temperture,Humidity},
{Outlook,Humidity,Windy},
{Outlook,Windy}.
```

The above reduct sets are used to generate new rules using naïve Bayesian classification algorithm.

Let us consider an example for generating new rules from the reduct set. Consider a reduct {temperature,humidity,windy} and classify the tuple X=(temperature=cool,humidity=high,windy=weak), which is not in the original database.To compute

P (D) probability for a decision attribute i.e. play ball.

P(playball=no)=39/78=0.5

P(playball=yes)=38/78=0.48

To compute P(C|D).Compute the following conditional attributes probabilities.

P(temperature=cool|playball=yes)=11/38=0.28,
P(temperature=cool|playball=no)=11/39=0.28,
P(humidity=high|playball=yes)=10/38=0.26,
P(humidity=high|playball=no)=29/39=0.74,
P(windy=weak|playball=yes)=21/38=0.552,
P(windy=weak|playball=no)=21/39=0.538.

Using the above probabilities, we obtain

P(X|playball=yes)=P(temperature=cool|playball=yes)*P(humidity=high|playball=yes)*P(windy=weak|playball=yes).
= 0.28*0.26*0.552
= 0.04*0.48(P(playball=yes))=0.0192

Similarly,

P(X|playball=no)=P(temperature=cool|playball=no)*P(humidity=high|playball=no)*P(windy=weak|playball=no).
= 0.28*0.74*0.538
= 0.05*0.5(P(playball=no))=0.0557

Therefore, the naïve Bayesian classification predicts playball = no for tuple X.

Let us consider reduct sets obtained from the Horizontal partitioning databases by using data cleaning algorithm.

{Temperature,Humidity,Windy},
{Humidity,Windy},
{Windy}

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

{Humidity},
{Temperature,Windy},
{Temperature},
{Temperature,Humidity},
{Outlook, Humidity,Windy},
{Outlook,Windy},
{Outlook},
{Outlook,Humidty},
{Outlook,Temperature},
{Outlook,Temperature,Humidity}.

The above reduct sets are used to generate new rules using naïve Bayesian classification algorithm.

Let us consider the above example for generating new rules from the reduct set. Consider a reduct {temperature,humidity,windy} and classify the tuple X=(temperature=cool, humidity=high,windy=weak), which is not in the original database. To compute

P (D) probability for a decision attribute i.e. play ball.

P(playball=no)=15/46=0.326

P(playball=yes)=31/46=0.673

To compute P(C|D).Compute the following conditional attributes probabilities.

P(temperature=cool|playball=yes)=8/31=0.258,
P(temperature=cool|playball=no)=4/15=0.266,
P(humidity=high|playball=yes)=9/31=0.29,
P(humidity=high|playball=no)=11/15=0.73,
P(windy=weak|playball=yes)=25/31=0.806,
P(windy=weak|playball=no)=4/15=0.266

Using the above probabilities, we obtain P(X|playball=yes)=P(temperature=cool|playball=yes)*P(humidity=high|playball=yes)*P(windy=weak|playball=yes).

= 0.258*0.29*0.806

= 0.06*0.326(P(playball=yes))=0.03

Similarly,

P(X|playball=no)=P(temperature=cool|playball=no)*P(humidity=high|playball=no)*P(windy=weak|playball=no).

= 0.266*0.73*0.266

= 0.05*0.67(P(playball=no))=0.01

Therefore, the naïve Bayesian classification predicts playball = yes for tuple X.

The above example generates two different rules for a given tuple. These rules are generated from the reduct sets of the uncleaned databases and cleaned databases. The below example shows two different values of a decision attribute for the same tuple.

X=(temperature=cool,humidity=high,windy=weak)

Playball = no,in the case of uncleaned database.

X=(temperature=cool,humidity=high,windy=weak)

playball=yes,in the case of cleaned database.

This shows that it is essential to apply data cleaning operation on the databases before calculating the reduct sets from the original databases. The attribute reduction algorithm is used to generate the reduct sets for cleaned databases and uncleaned databases. The reduct sets which are generated from uncleaned databases are different from cleaned databases. The rules which are generated from the reduct sets of the uncleaned databases are not accurate, because the original database contains the noisy data, redundant data and inconsistent data. But the rules which are generated from the reduct set of the cleaned databases are definitely accurate, as it does not contain noisy data, redundant data and inconsistent data. The rules generated by the uncleaned databases are not useful to the researchers for their research activities. So, data cleaning is an essential operation on the databases before generating the rules from the reduct set.

Time Complexity Graphs:

This paper presents the graphs comparing the time complexity for finding Reduct sets before cleaning databases and after cleaning databases.

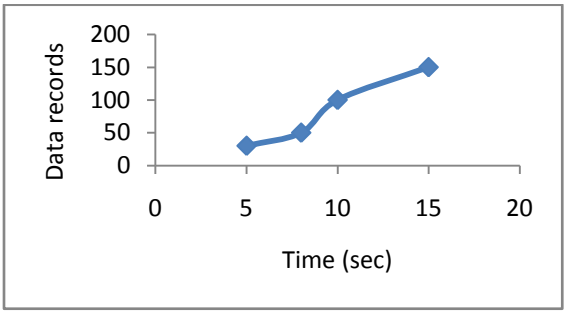

Figure1: Without cleaning Time complexity

Figure1 shows the execution time for finding reduct sets before removing noisy data, duplicate data and inconsistent data in the original databases. If the database contains noisy and inconsistent data, it takes more time to find the reduct sets.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

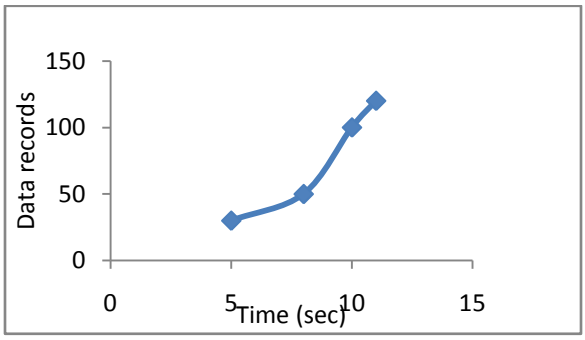

Figure2: with cleaning data time complexity

Figure2 shows the execution time for finding reduct sets after removing noisy data, duplicate data and inconsistent data in the original databases. After removing noisy, inconsistent and redundant data it takes less time to find out the reduct sets.

V. CONCLUSION

This paper describes a method for finding a roughset concept called reduct set from partitioned databases using secure multiparty technique. From this reduct set, rules are generated using Naïve Bayesian algorithm. This paper also describes attribute reduction technique for horizontal partitioning of databases. This paper proposes an enhanced transitive closure algorithm to remove noisy, duplicate and inconsistent data from the original micro data. In the present work only two different databases are considered to generate new rules, but in future there is a necessity to consider the multiple databases to generate new rules.

REFERENCES

[1]. Mingquan Ye ; Xuegang Hu ; Changrong Wu,"privacy preserving Attribute Reduction for Horizontaly Partitioned Data" ,Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on
[2]. Mingquan Ye ; Xuegang Hu ; Changrong Wu , **"**Privacy Preserving Attribute Reduction for Vertically Partitioned Data", 2010 International Conference on Artificial Intelligence and Computational Intelligence.
[3].Z. Pawlak,"Rough sets", Journal of Computer and Information Science,11(5), pp. 341-356,1982.
[4]. X. Lin, C. Clifton, M. Zhu, "Privacy-preserving clustering with distributed EM mixture modeling", Knowledge and Information Systems,8(1), pp. 68-81, 2005.
[5]. Z. Pawlak, Rough set: Theoretical Aspects and Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.
[6]. Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma," HADCLEAN:A Hyboritrid Approach to Data Cleaning in Data Warehouses",Department of Computer Science & Information Systems Birla Institute of Technology & Science.