Field Level Security of the Sensitive Data in Large Datasets

K. Shirisha, K. Haritha

Abstract: ***The process of deriving useful and knowledgeable information from enormous quantity of data is Data Mining. During mining procedures, handling of the sensitive data has become important to protect data against illegal attacks and malicious access either during transmission or at rest. Association rule algorithm is one of the rule extraction techniques. The rules determined are either to be transferred over the public networks or to be rested for further use.The main objective of the Field Level Security of the Sensitive Data in Large Datasets is to extract the strong association rules from the large data sets and the outcomes are crafted to conceal the sensitive data. The datasets and the association rules involving the attributes with relationships and dependencies are modified through several approaches and to see that no sensitive association rule is derived from it[1]. Privacy preservation of the sensitive association rules in large datasets is to provide secrecy for the sensitive data. Presently, it has become quite important to safeguard the privacy of the users' personal data from unauthorized persons. The usage of association rules in voluminous datasets has emerged to be advantageous to organizations [2]. In this paper, we present a novel approach which is applied for hiding sensitive association rules by utilizing the techniques of compression, encryption method ology on the original dataset, providing dataset with better immunity.***

Keywords: ***Association rules, Data mining, Privacy preservation,Sensitive data.***

I. INTRODUCTION

The method of extracting helpful knowledge from massive amounts of facts is Data mining. Knowledge had to be crafted in such a way that the data is not extracted using any of the techniques of data processing. For extracting out the knowledge and the interesting patterns, a number of procedures are used, depending upon the domain of the application and the kind and granularity of the data contained in the raw form. Unlike the conventional way, the data for mining is staged at multiple locations geographically and remain in the distributed architectural model. The mined information from one stage is to be transferred to the other locations for integration and subsequent mining procedures.

While handling sensitive data, it becomes important tosafeguard knowledge against unauthorized access while being transferred. The likely result of the this sorry state of affairs ends up in the analysis of sensitive data concealing in information [4] for malicious purposes. The increased facility of storing the personal data of the users using advanced algorithms has led to the necessity of privacy conservation. Varieties of techniques are advised in recent years so as to perform privacy conserving data processing. Sensitive data is identified and hidden using many algorithms, namely, heuristic algorithm, borderbased approach, etc. These algorithms provide only a single level protection from hackers.[2]. In this paper, the Section 2 discusses about the Association Rule mining methodology and the related popular approaches and tools to address the same. In Section 3, the system model for the Field Level Security of the Sensitive Data in Large Datasets is discussed. In Section 4, the algorithmic procedures are elaborated for implementing the security measures on the sensitive data to be transmitted over the public networks. The experimental results are detailed in the next section. In the section 6, the conclusions and future scope are mentioned.

II. ASSOCIATION RULE MINING AND APPROACHES

Association Rule Mining (ARM) is a method to produce frequently occurring patterns and to determine the correlations or causal structures among the data itemsets in data repositories bearing the variety of data structures to represent the real- world data. It is used to discover unique and useful pattern from enormous volume of data. R.Agarwal[3] in 1993 has found the technique of data mining.

Suppose itemsets are $P=\{p_1, p_2,...., p_m\}$ and the transaction sets are $M=\{m_1, m_2,....,m_l\}$. Unique id is assigned for every transaction. Any transaction say $m_j \mid j=[1\,..\,l]$ contain the combination of one or more items $p_i \mid i=[1..m]$ from the itemset P.

Let the implication $A \Rightarrow B$ be the association rule, where A *and* B are the subsets of itemset in P and $A \cap B = \Phi$. In the association rule, $A \Rightarrow B$, the A is called as the antecedent and B is the consequent. Although a large number of above mentioned implications

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author

K.Shirisha*, Professor, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana State, India. Email: shirisha49@yahoo.com

Haritha Kunta, Associate Success Agent, Salesforce.com India Pvt. Ltd. Hyderabad, India. Email: haritha.k9909@gmail.com

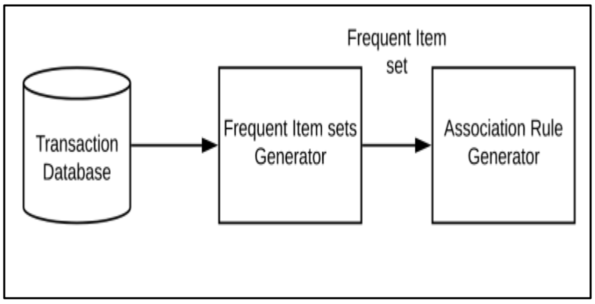


Fig.1. Association rule mining

emerge, only few of them remain really interesting and potentially of use. The interestingness measures that are considered for determining the potentially significant rules are the support and the confidence parameters.The criteria for establishing the rule as potentially of use and novel are that the implication has to surpass minimum support threshold, $\min\_\sup\_threshold$, as well as the minimum confidence threshold, $\min\_conf\_threshold$. The prerequisite to be observed for declaring the implication as frequent is the condition:

$$Support \quad (A \Rightarrow B) \geq \min\_\sup\_threshold$$

where the $\min\_\sup\_threshold$ is the proportion which is defined as:[4] $Support \quad (A \Rightarrow B) = |A \cap B| / |N|$.

The $|A \cap B|$ is the number of total transactions in database containing the itemsets A and B both.

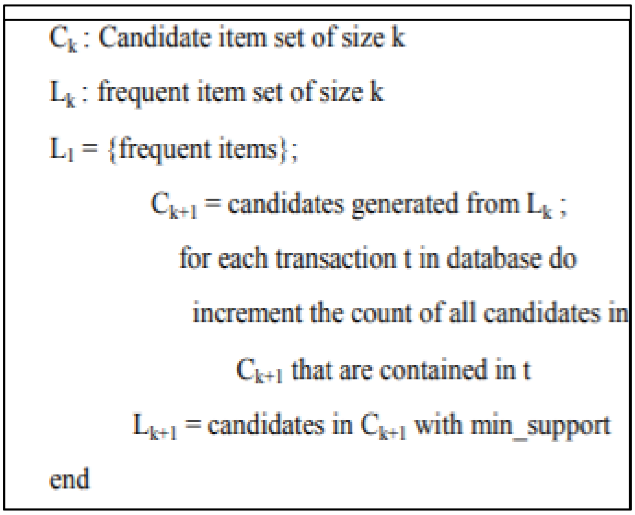
The second interestingness measure to pronounce the implication of the frequent itemset as the strong association rule is the condition:

$$Confidence\,(A \Rightarrow B) \geq \min\_conf\_threshold$$

where $Confidence\ (A \Rightarrow B) = |A \cap B| / |A|$. The $|A|$ denote the number of total transactions contained in the database N that have the itemset A. A rule $A \Rightarrow B$ is said to be a strong association rule if both the conditions are exercised.

A. Apriori Algorithm

The Apriori algorithm is the popular technique to find the frequent itemsets from the transactional dataset, the first step towards finding the strong association rules. This algorithm works by finding the repeated itemsets with respect to the $\min\_\sup\_threshold$, from the N and thereafter increasing the length of the itemsets by one in every subsequent iteration. All the itemsets which are repeated be (k-1) has k-itemsets that are frequent are superset. The following shows the generation of candidate using Apriori.[5]



```
$C_k$ : Candidate item set of size k
$L_k$ : frequent item set of size k
$L_1$ = {frequent items};
        $C_{k+1}$ = candidates generated from $L_k$ ;
          for each transaction t in database do
            increment the count of all candidates in
                $C_{k+1}$ that are contained in t
        $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
end
```

Fig.2.Apriori Algorithm

Weka is a standard tool which is used to perform algorithms based on machine learning techniques. This may include data cleansing, data transformation etc.[6]

III. FIELD LEVEL SECURITY OF THE SENSITIVE DATA IN LARGE DATASETS

The proposed system for imposing Field Level Security of the Sensitive Data in Large Datasets uses an algorithm that conceal the sensitive association rules by majorly following the two step process:

a. Encryption of sensitive data.
b. Compression of the dataset.

Initially, the sensitive data is identified from the original data and the association rules are generated. The data emerged in the association rules are mapped to the fields identified as sensitive are further encrypted to conceal their confidentiality. Furthermore, the additional step of compression is applied to lower the load of transmission among the low configured devices and networks. This compressed file is shared through open network. This makes attacker difficult to extract the original data, as it is in the encrypted and compressed form.

IV. SYSTEM MODELING OF THE FIELD LEVEL SECURITY OF THE SENSITIVE DATA

The architectural design of the system is modeled by subdividing the system into the following tasks:

a. Data pre-processing
b. Identification of the sensitive data
c. Encryption of the sensitive data
d. Conversion of information representation from numerical to nominal
e. Generation of strong association rules
f. Compression of the encrypted association rules.

The series of tasks is diagrammatically presented in the Fig.3. The outcome is the compressed file which is to be shared with the authorized receiver over the public channel.

The dataset for the illustrative purpose is chosen from the medicarewebsite[7]. This dataset is initially preprocessed using the methods in the WEKA tool. The sensitive data is identified from the preprocessed data. Sensitive data columns are encrypted using the encryption technique.[8] The Apriori algorithm used for the generation of association rules from WEKA is following a precondition for the format of the inputting data to be only in nominal form. Conversion of the data representation from the dissimilar form to the nominal form is performed.

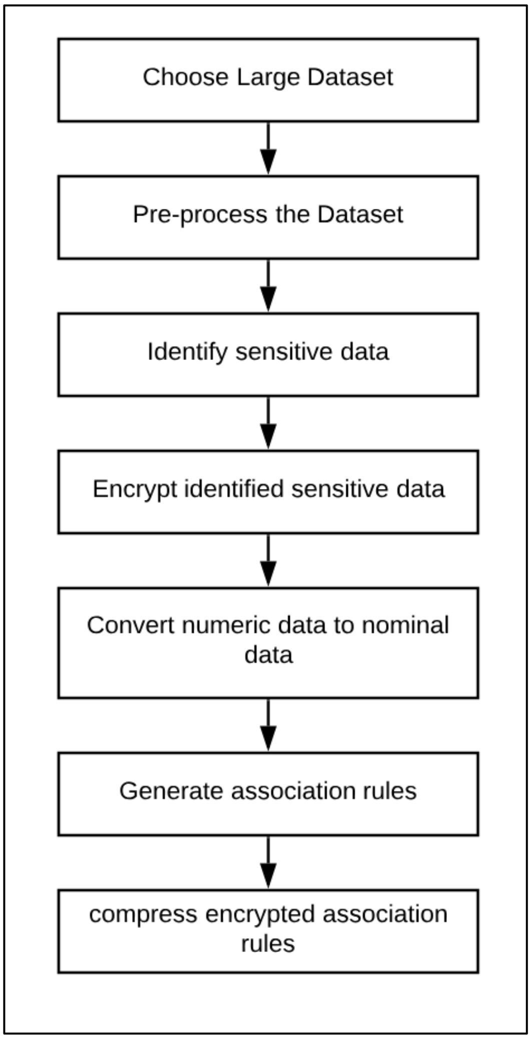

Choose Large Dataset
Pre-process the Dataset
Identify sensitive data
Encrypt identified sensitive data
Convert numeric data to nominal data
Generate association rules
compress encrypted association rules

Fig. 3. Architectural Design

The compression is applied so as to further strengthen the confidentiality feature of the association rules to be shared across.

V. ALGORITHMIC AND EXPERIMENTAL SET-UP

The following algorithms are implemented to achieve the preprocessing, conversions, encryption, generation of strong rules, and compression techniques. The MD5 algorithm is used to encrypt the sensitive information for sharing.

Algorithm Integration

```
Begin
    1. Create object for replacemissingvalues
    2. Create object for hash_col_en
    3. Create object for csv2arff
    4. Create object for num2nom
    5. Create object for apriori
  6. Create object for compression
End
```

Algorithm ReplaceMissingValues[9]

```
Begin
//Inorder to replace any null values present in the source file
//with maximum value of the field, we use a method from
//WEKA and save it as CSV file.
    1. Acquires the source file path and reads the file.
    2. Applies ReplaceMissingValues() method from
       WEKA //replaces the null values with the
       maximum values.
    3. Save it as a file in CSV format by giving a
       destination.
End
```

Algorithm hash_col_en[10]

```
Begin
//The separated sensitive fields are encrypted and amended
//to the original file.
    1. Acquires the source file path and reads the file.
    2. Splits the sensitive columns and saves it as text.
    3. Create MessageDigest instance for MD5
    4. Add input text bytes to digest.
    5. Convert these bytes to hexadecimal bytes.
    6. Get complete hashed text in hex format.
    7. Replace the original value in the file with obtained
       value.
End
```

Algorithm csv2arff[11]

```
Begin
//Convert CSV file format to Arff format
    1. Load a CSV file using CSVLoader from WEKA.
    2. Save the file as Arff using ArffSaver from WEKA.
End
```

Algorithm num2nom[12]

```
Begin
//To convert numeric data to nominal data.
    1. Load the Arff file which was saved earlier.
    2. Call NumericToNominal() //converts numeric data
       to nominal data.
    3. Save the resultant file as Arff using ArffSaver.
End
```

Algorithm apriori[13]

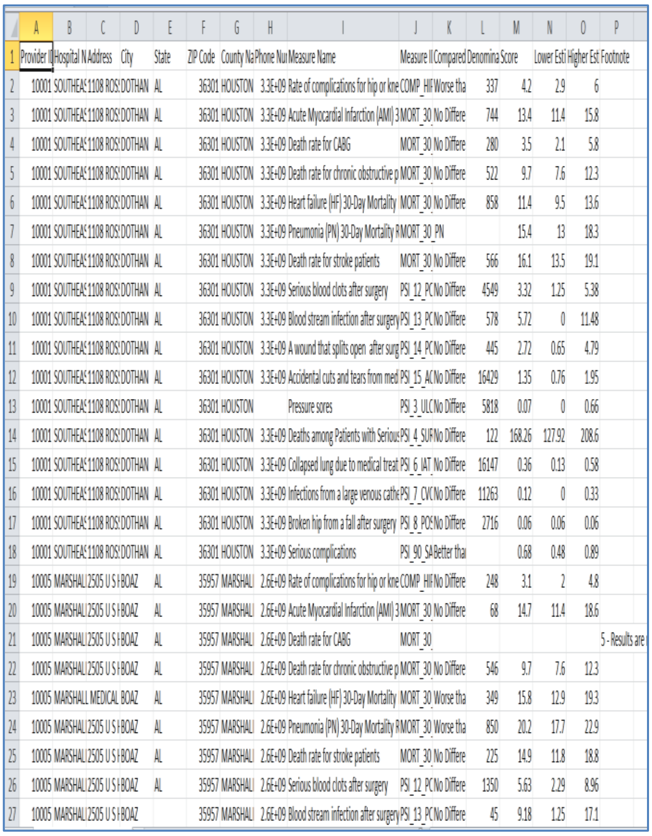
```
Begin
//The associations are built by calling the Apriori
    1. Load the Arff file which was saved earlier.
    2. Call Apriori() to buildAssociations.
    3. Save it to a file to view.
End
```

Algorithm compression[14]

```
Begin
//Write to a file.
    1. Load the file which was saved earlier as an
       InputStream.
    2. Use DeflaterOutputStream(destination file)
    3. Write it to a file to view.
End
```

VI. EXPERIMENTAL RESULTS

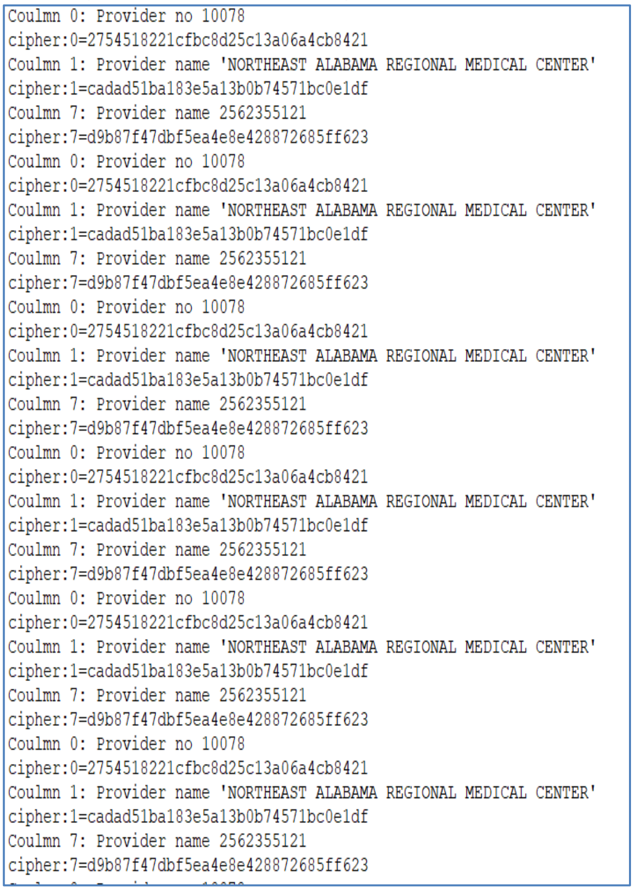
The partial input, taken from online source[7], represented with the screenshot in the Fig.4 is subjected to the mentioned series of the cleansing, conversions, encryption and compression techniques using the algorithmic implementation as described in the Section 5.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Provider II	Hospital N	Address	City	State	ZIP Code	County Na	Phone Nui	Measure Name	Measure II	Compared	Denomina	Score	Lower Esti	Higher Est	Footnote
2	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Rate of complications for hip or kne	COMP_HIF	Worse tha	337	4.2	2.9	6	
3	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Acute Myocardial Infarction (AMI) 3	MORT_30	No Differe	744	13.4	11.4	15.8	
4	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Death rate for CABG	MORT_30	No Differe	280	3.5	2.1	5.8	
5	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Death rate for chronic obstructive p	MORT_30	No Differe	522	9.7	7.6	12.3	
6	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Heart failure (HF) 30-Day Mortality	MORT_30	No Differe	858	11.4	9.5	13.6	
7	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Pneumonia (PN) 30-Day Mortality R	MORT_30_PN			15.4	13	18.3	
8	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Death rate for stroke patients	MORT_30	No Differe	566	16.1	13.5	19.1	
9	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Serious blood clots after surgery	PSI_12_PC	No Differe	4549	3.32	1.25	5.38	
10	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Blood stream infection after surgery	PSI_13_PC	No Differe	578	5.72	0	11.48	
11	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	A wound that splits open after surg	PSI_14_PC	No Differe	445	2.72	0.65	4.79	
12	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Accidental cuts and tears from med	PSI_15_AC	No Differe	16429	1.35	0.76	1.95	
13	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON		Pressure sores	PSI_3_ULC	No Differe	5818	0.07	0	0.66	
14	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Deaths among Patients with Serious	PSI_4_SUF	No Differe	122	168.26	127.92	208.6	
15	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Collapsed lung due to medical treat	PSI_6_IAT	No Differe	16147	0.36	0.13	0.58	
16	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Infections from a large venous cathe	PSI_7_CVC	No Differe	11263	0.12	0	0.33	
17	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Broken hip from a fall after surgery	PSI_8_POS	No Differe	2716	0.06	0.06	0.06	
18	10001	SOUTHEAS	1108 ROS	DOTHAN	AL	36301	HOUSTON	3.3E+09	Serious complications	PSI_90_SA	Better thai		0.68	0.48	0.89	
19	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Rate of complications for hip or kne	COMP_HIF	No Differe	248	3.1	2	4.8	
20	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Acute Myocardial Infarction (AMI) 3	MORT_30	No Differe	68	14.7	11.4	18.6	
21	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Death rate for CABG	MORT_30						5 - Results are
22	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Death rate for chronic obstructive p	MORT_30	No Differe	546	9.7	7.6	12.3	
23	10005	MARSHALL MEDICAL		BOAZ	AL	35957	MARSHALI	2.6E+09	Heart failure (HF) 30-Day Mortality	MORT_30	Worse tha	349	15.8	12.9	19.3	
24	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Pneumonia (PN) 30-Day Mortality R	MORT_30	Worse tha	850	20.2	17.7	22.9	
25	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Death rate for stroke patients	MORT_30	No Differe	225	14.9	11.8	18.8	
26	10005	MARSHALI	2505 U S H	BOAZ	AL	35957	MARSHALI	2.6E+09	Serious blood clots after surgery	PSI_12_PC	No Differe	1350	5.63	2.29	8.96	
27	10005	MARSHALI	2505 U S H	BOAZ		35957	MARSHALI	2.6E+09	Blood stream infection after surgery	PSI_13_PC	No Differe	45	9.18	1.25	17.1	

Fig.4. Screenshot of partial dataset

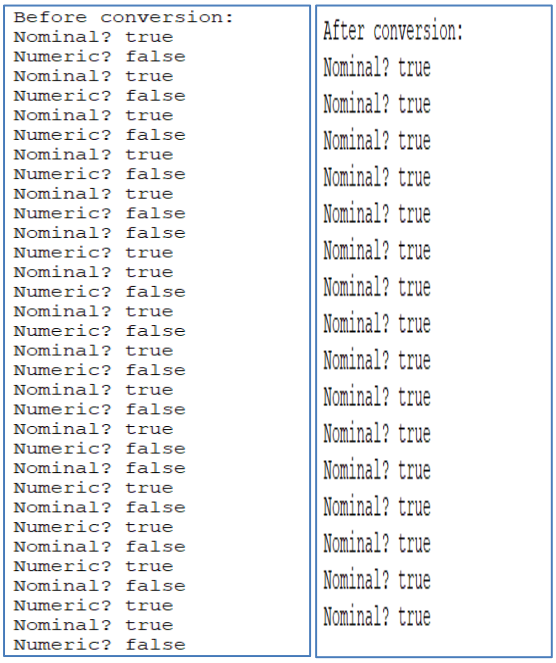
Firstly the data is preprocessed and the sensitive columns are encrypted as shown in the figure below.



```
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
Coulmn 0: Provider no 10078
cipher:0=2754518221cfbc8d25c13a06a4cb8421
Coulmn 1: Provider name 'NORTHEAST ALABAMA REGIONAL MEDICAL CENTER'
cipher:1=cadad51ba183e5a13b0b74571bc0e1df
Coulmn 7: Provider name 2562355121
cipher:7=d9b87f47dbf5ea4e8e428872685ff623
```

Fig.5.Screenshot of the encrypted columns

Then it is checked whether the encrypted file has any numeric data, if detected it is converted to nominal data. This is done so as to assure that the file given to apriori algorithm for generation of association rules takes only nominal data.

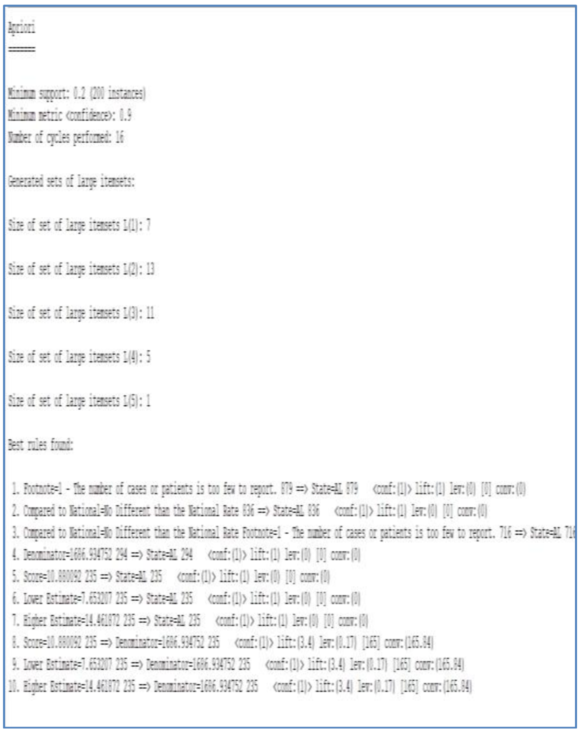


```
Before conversion:
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? false
Numeric? true
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? true
Numeric? false
Nominal? false
Numeric? true
Nominal? false
Numeric? true
Nominal? false
Numeric? true
Nominal? false
Numeric? true
Nominal? true
Numeric? false
```

```
After conversion:
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
Nominal? true
```

Fig.6. Conversion of numeric data to nominal data

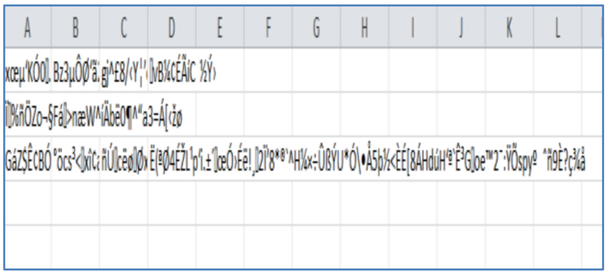
Then the best rules of association are generated using Apriori algorithm.



```
Apriori
=======

Minimum support: 0.2 (200 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 13

Size of set of large itemsets L(3): 11

Size of set of large itemsets L(4): 5

Size of set of large itemsets L(5): 1

Best rules found:

 1. Footnote=1 - The number of cases or patients is too few to report. 879 ==> State=AL 879    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. Compared to National=No Different than the National Rate 836 ==> State=AL 836    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. Compared to National=No Different than the National Rate Footnote=1 - The number of cases or patients is too few to report. 716 ==> State=AL 716
 4. Denominator=1686.934752 294 ==> State=AL 294    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. Score=10.880092 235 ==> State=AL 235    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. Lower Estimate=7.653207 235 ==> State=AL 235    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. Higher Estimate=14.461872 235 ==> State=AL 235    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. Score=10.880092 235 ==> Denominator=1686.934752 235    <conf:(1)> lift:(3.4) lev:(0.17) [165] conv:(165.84)
 9. Lower Estimate=7.653207 235 ==> Denominator=1686.934752 235    <conf:(1)> lift:(3.4) lev:(0.17) [165] conv:(165.84)
10. Higher Estimate=14.461872 235 ==> Denominator=1686.934752 235    <conf:(1)> lift:(3.4) lev:(0.17) [165] conv:(165.84)
```

Fig.7.Screenshot of the generated Apriori rules

The generated rules are then compressed and sent to the receiver via open network. This encryption followed by compression technique makes the hacker difficult to hack.



xœµ'KÓ0‖ Bz3µÔØ'ã; gj^£8/‹Y¦'‹‖vB¾CÉÃiC ½Ý›

‖‖%ñÖZo–§Fá‖›næW^iÄbë0¶^"a3=Á[‹žø

GáZŞÊCBÓ°öcs³<‖xîC; ñÚ‖cëø‖Øx Ë(ªØ4ÉŽL¹p'í.±'‖œÓ›Éë!‚‖2Ï'8*®'^H¼x÷ÛßÝU*Ó\•Å5þ½<ÈÉ[8ÁHdúH'ª`Ê³G‖oe™2¨:ŸÕspyº ^ñ9È?ç¾â

Fig.8.Screenshot of the compressed file

VII. CONCLUSION AND FUTURE SCOPE

Privacy preservation of the sensitive association rules in large datasets is achieved to provide secrecy for the sensitive data. The system proposed makes the data more secure and display only required information excluding the sensitive data for the third party which makes it difficult to hack. Every system may not meet all the requirements of the user as the requirements of the user changes as he uses the system.

Further improvisations to this system are:

1. Security can be improved using latest encryption and compression techniques based on the future security issues.
2. The minimum support as well as the minimum confidence can be made customizable to the user[15]

REFERENCES

1. Chandrima R Ghosh1, Jasmine Jha2, “An Enhanced Association Rule Mining To Find Sensitive Patterns And Hide Them For Privacy Preservation”, IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-3 2016
2. D. Bhanu, P. Balasubramanie, “Predictive modeling of inter-transaction association rules – a business perspective.”, International Journal of Computer Science and Applications,Vol. 5, No. 4, pp57– 69 , 2008.
3. R. Agrawal, T.Imielinski, and A. Swami, R.Srikant, “Mining association rules between sets of items in large databases,”In Proceedings of ACM SIGMOD International Conference on Management of Data, Washington, DC, pp. 207-216,May 1993.
4. KenampreetKaur, MeenakshiBansal ”A Review on various techniques of hiding Association rules in Privacy Preservation Data Mining” IJECS Volume 4 Issue 6 June, 2015 Page No.12947-12951.
5. S.Narmadha, S.Vijayarani, ”Protecting Sensitive Association Rules in Privacy Preserving Data Mining using Genetic Algorithms.”, International Journal of Computer Applications (0975 – 8887),Volume 33– No.7, November 2011.
6. https://www.cs.waikato.ac.nz/~ml/weka/
7. https://data.medicare.gov/data/hospital-compare
8. https://examples.javacodegeeks.com/category/core-java/crypto/
9. http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/ReplaceMissingValues.html
10. https://howtodoinjava.com/security/how-to-generate-secure-password-hash-md5-sha-pbkdf2-bcrypt-examples/
11. https://stackoverflow.com/questions/10341701/convert-csv-to-arff-using-weka
12. http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/NumericToNominal.html\
13. http://www.cs.tufts.edu/~ablumer/weka/doc/weka.associations.Apriori.html#Apriori()
14. https://www.geeksforgeeks.org/compressing-and-decompressing-files-in-java/
15. https://www.researchgate.net/publication/233754781_Support_vs_Confidence_in_Association_Rule_Algorithms/
Fig.1.Association rule mining Source: Adapted from [4]
Fig.2. Apriori Algorithm source: Adapted from [5]

AUTHORS PROFILE

Dr. Shirisha Kakarla, Professor in the Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology. She pursued her B.E. in Computer Science, M.Tech.in Software Engineering and Ph.D. programme in area of Information Security. She has more than eighteen years of teaching experience. Her research interests include information and cyber security, data mining, database security, machine learning and data structures. She is a life member of ISTE. She published more than twenty six research articles in the leading international journals with a number of them in Scopus. She contributed a chapter in the Elsevier journal and also represented India in SAARC International Conference in 2018 with a scholarly article.

Haritha Kunta, is currently working as an Associate Success Agent within salesforce.com India Pvt. Ltd. She pursued Bachelor’s of Technology in Computer Science and Engineering at Sreenidhi Institute of Science and Technology. Her strong interest in Computer Science has enabled her to develop skills in Data Structures, Object Oriented Programming, Data Analysis and Algorithms, Database Management, Information Security. Besides the undergraduate program, she completed few courses on Game Development in C, Android Development, and Internet of things. The project on Internet of Things enabled her to enhance her programming skills and learn about this new technology.