I.J. Information Engineering and Electronic Business, **2018, 3, 45-50**
Published Online May 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2018.03.06

A Novel Algorithm for Association Rule Hiding

T.Satyanarayana Murthy[1], **N.P.Gopalan**[2]
Department of Computer Applications, National Institute of Technology, Tiruchirapalli
Email: [1]murthyteki@gmail.com,[2]npgopalan@gmail.com

Received: 28 August 2017; Accepted: 08 January 2018; Published: 08 May 2018

***Abstract*—**Current days privacy concern about an individual, an organization and social media etc. plays a vital role. Online business deals with millions of transactions daily, these transactions may leads to privacy issues. Association rule hiding is a solution to these privacy issue, which focuses on hiding the sensitive information produces from online departmental stores ,face book datasets etc..These techniques are used to identify the sensitive rules and provide the privacy to the sensitive rules, so that results the lost rules and ghost rules. Algorithms developed so far are lack in achieving the better outcomes. This paper propose two novel algorithm that uses the properties from genetic algorithm and water marking algorithm for better way of hiding the sensitive association rules.

***Index Terms*—**Privacy, privacy preserving, lost rules, ghost rules

I. Introduction

Online business to deal with millions of transactions. Privacy for the sensitive information is a major challenging task. Association rule mining [1,2,3] a major technique for market basket analysis where processing the transactions and generate the association among these transactions. These association are based on the parameters like support and confidence. support determines the occurence of the item appears in the transactional dataset, where as confidence determines the strength of the rule. These rules are divided into SR and NSR based on the minimum support and maximum confidence parameters. This article focuses on hiding the sensitive rules. During this hiding process ,hiding failure is a major parameter determines the failure of hiding the sensitive association rules. In this paper, identify the privacy breach while hiding the sensitive association rules. Our goal is reduce the ghost rules and increasing the identification of lost rules. The objective of the article takes the dataset as an input and applies the Aprior rule miner for generating the associations rules among the dataset. Instead of Aprior Algorithm using the Association rule hiding algorithm generates a Sanitized Dataset. DSRRC approach cannot handle ,hiding association rules with multiple items in. The efficiency of MDSRRC algorithm can be poor, so it must require future enhancements. Genetic Algorithm [4,5] is an evolutionary algorithm evolves from the lives of the human beings. Genetic algorithm is made up of collection of individuals called chromosomes. These chromosomes are used to represents the population to develop a solution for the suitable problem. Each chromosome represented with a binary values either 0 or 1. The basic theory behind the genetic algorithm was the survival of fittest proposed by Darwin. The species that live longer can have more fitness leads to more survival less fitness leads to less survivals. The Genetic Algorithm Approach begins with random generation of individuals. These individuals collection together called as an population. During the genetic Algorithm process a new population replaces the old population in each iteration. The best chromosome in the population are chosen. The population will transform into future chromosomes basing on the fitness function. The operations that are performed are initialize the population, selection of chromosomes, calculate the fitness values based on the random values. Association rule mining used for analyze the transactional database for identifying strong rules. This was proposed by Agrawal and Cheung for transactional data-sets. The main purpose is finding the frequently used item sets and generate association rules on the dataset. Let a Transactional Database D consists of t1,t2,t3,t4,t5...tn where T is a collection of items like i1,i2,i3,...in. Support of X->Y determines the the ratio of the records which contains XUY with the D, where D equals number of transactions. Confidence gives the strength of the rules..

II. Literature Survey

Heuristic, Border, Exact, Reconstruction based and Cryptographic based techniques are used for hiding the sensitive association rules. V. Verykios team [6] mainly focused on hiding the sensitive rules using a cyclic algorithmic approach so that limiting disclosure of sensitive rules. Elena Dasseni, Vassilios S. Verkios,Ahmed K. Elmagarmid,Elisa Bernito [7] proposed methods based on confidence and support. V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni [8] proposed five algorithms to hide the sensitive knowledge. Stanley R. M. Oliveira, Osmar R. Zaiane [9] improved the balancing factor for achieving the better privacy results. ADSRCC and RRLR algorithm are used for hiding the sensitive rules was proposed by Komal Shah, Amit Thakkar,Amit Ganatra [10].ADSRRC algorithm was popular algorithm to hide the sensitive rules based on the strategy i.e support-based technique. RRLR algorithm uses confidence based technique for hiding the sensitive rules. RRLR algorithm performs more efficiently than ADSRCC. Damandiya,UdayPratap Rao [11] overcomes the limitations with ADSRCC and proposed a novel technique called MDSRRC algorithm based on the support reduction technique in 2013. Tzung-Pei Hong,Chun-Wei Lin,Kuo-Tung Yang,Shyue-Liang Wang [12] proposed a SIF-IDF algorithm for hiding the sensitive rules based on the support based strategy. This algorithm takes more running time for hiding the sensitive rules. To reduce the running time Chun-Wei Lin, Tzung-Pei Hong, and Hung-Chuan Hsu [13] proposed a HMAU Algorithm for hiding the sensitive rules based on the support. Narges Jamshidian, Ghalehsefidi,Mohammad Naderi Dehkordi [14] developed a hybrid algorithm based on the support and confidence based strategy. Belwal R., Varshney J, Khan S [15] hiding the sensitive rules by reducing the support and confidence. C. N. Modi, U. P. Rao, and D. R. Patel [16,17] proposed an algorithm based on clustering to reduce the side effects on sanitized database.DSRRC approach fails to hide multiple association rules. B.Kesava Murthy, Asad M.Khan [18] uses Genetic Algorithm. Chun-Wei, Jerry Lin [19] proposed a sanitization approach for hiding sensitive item sets based on PSO. Mahtab Hossein Afshari [20] proposed an association rules hiding technique using cuckoo optimization Algorithm. T.Satyanarayana Murthy [21,22] proposed novel algorithms for privacy preserving data mining. N.P.Gopalan, T.Satyanarayana Murthy [23] uses CRO for privacy preserving.N.P.Gopalan, T.Satyanarayana Murthy, Yalla Venkateswarlu [24] uses a novel approach for hiding critical transactions using Un-realization Approach.T.Satyanarayana Murthy et al [25] proposed an efficient method for hiding association rules with additional parameter metrics.

III. Problem Statement

A. Terminology Used

Table 1. Terminology

D	Original Transactional Dataset
SD	Sanitized Database
R	Association Rules
SR	Sensitive Rules SR subset of R
NSR	Non Sensitive rules
t1,t2,t3,t4.....tn	Set of transactions.
i1,i2,i3,i4...in	Set of items
MST	Support Threshold
MCT	Confidence Threshold

B. Problem Definition

Database D a collection of transactions t1,t2,t3……tn where transaction is a collection of items i1,i2,i3……in. Generate the rules using an Association Rule miner on the database D for mining the rules R. These rules are divided into sensitive rules and non-sensitive rules based on the support and confidence. Now proposed a Novel algorithm uses watermarking mathematical model for sanitizing the database D into SD. SD given as an input to the rule miner results non disclosure of sensitive rules.

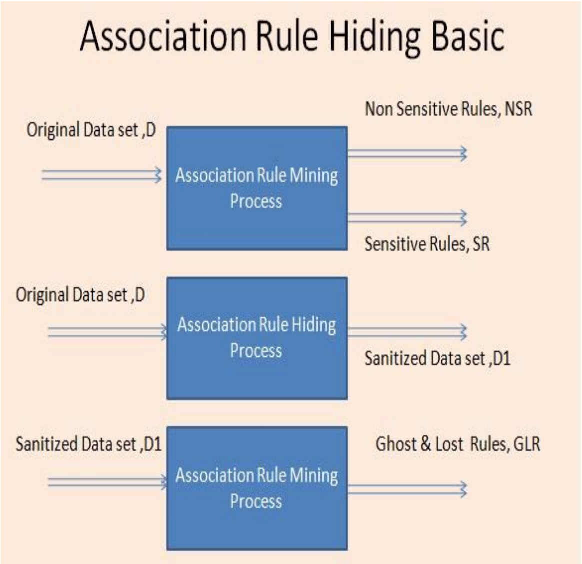

Fig.1. Association Rule Hiding Basic.

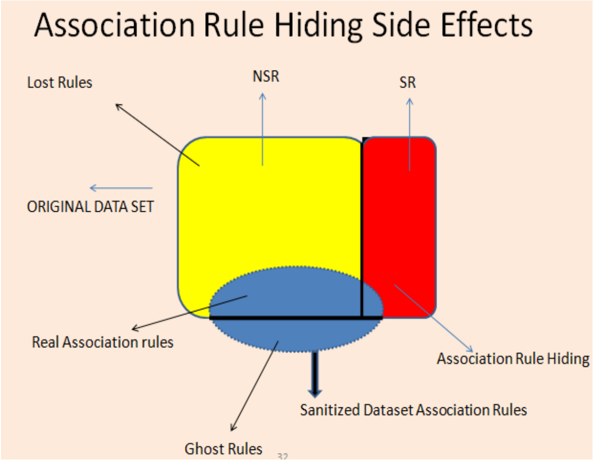

Fig.2. Association Rule Hiding side effects.

IV. CPT Alorithm

The CPT [26] scheme uses a matrix in the binary form and a weight matrix as secret key.CPT algorithms mainly uses XOR operations. Weight matrix mainly used to increases the hiding data.F a binary image .Embed the data on F results modified binary image. It divides into m*n blocks. K a secret key shared between the sender and the receiver. W a weight matrix mainly used to increase the hiding capacity represents the number of bits embed on F.B a critical information embed in F.

V. Proposed Alorithms

Algorithm MAXWT1:

```
Input : Original Transactional Dataset D , MST , MCT, support (S), confidence (C),Transactional Array
        A,MAX_SUP,MAX_CONF
Output: Transactional Dataset D with sensitivity i.e reduces ghost rules and hike in lost rules.
    1. Given a Transactional Dataset D, D={t1,t2,t3....tn}.
    2. Apply Aprior Algorithm for generating the association rules.
    3. For all the rules R, calculate the Support s, Confidence c
    4. Find the MAX_SUP & MAX_CONF constant values.
    5. Identify the sensitive rules SR based on the MAX_SUP and MAX_CONF.
    6. Generate Sanitized Dataset using water_mark(D).
    7. Apply Apriori Algorithm on SD so that it retrieves rules that minimizes the ghost rules and increases the identification of lost rules..
    8. Repeat step2.
water_mark(Dataset D){
        original dataset D
        convert to Binary Dataset BD
        initialize key matrix K;
        initialize weight matrix W;
        initialize identity matrix id1,id2,id3;
        Given a numerical dataset D converted into
                binary data format.
        Convert the binary data set BD partioned into
                matrix of size m*n.
        Given a random block R of size m*n.
        Apply key matrix
        Let apply the weights W of size m*n where each
                cell value replaces at positions
                $10,10^{1,}10^{2}......10^{n}$ etc...
        do embedding process for hiding the data.
}
```

Algorithm MAXWT2 :

```
Input : Original Transactional Dataset D , MST , MCT, support (S), confidence (C),Transactional Array
        A,MAX_SUP,MAX_CONF
Output: Transactional Dataset D with sensitivity i.e reduces ghost rules and hike in lost rules.
    1. Given a Transactional Dataset D, D={t1,t2,t3....tn}.
    2. Apply Aprior Algorithm for generating the association rules.
    3. For all the rules R, calculate the Support s, Confidence c
    4. Find the MAX_SUP & MAX_CONF constant values.
    5. Identify the sensitive rules SR based on the MAX_SUP and MAX_CONF.
    6. Generate Sanitized Dataset using unrealize(D).
    7. Apply Apriori Algorithm on SD so that it retrieves rules that minimizes the ghost rules and increases the identification of lost rules..
    8. Repeat step2.
unrealize(Dataset D){
    If $T_S$ ==NULL reutrn UD
    t← count( $T_S$)
    UD=D;
    for (i=1; i<=t; && i%2==0;i++)
    update UD alternatively.
    UD----UD+D
    $T^P$←$T^P$– {t}
    t'← the most frequent dataset in $T^P$
    return unrealized –training set
}
```

VI. Experimental Results

Experiments were conducted on Core i5 Processor with 16GB of RAM running a Windows operating system and uses WEKA Data mining software. Experiments are conducted on Adult dataset, made up of 32561 tuples. Among that removes the missing value tuples, so that the left over tuples are 30722.This dataset consists of 14 attributes like salary, age, sex and marital status etc... Association rules are generated by using Weka tool and a open source python program. Python program accepts support value and confidence values as input and produces results. Experimental Results shows that MAXWT1, MAXWT2 approaches are compared with the DSRRC and MDSRRC approaches. The parameters used are Hiding Failure, Misses Costs, Artificial patterns, Dissimilarity, Side Effect Factor ,Lost rule recovery and Ghost rule generation. During the sanitization process failure of hiding the rules are determined by the parameter named Hiding Failure. Misses cost determines the non-sensitive data that hidden during the hiding process. Artificial patterns results the discovered patterns that are artificial. Dissimilarity measures the variation among the original and sanitized datasets. Side effect factor gives the amount of non sensitive rules removed. Lost Rule Recovery determines the percentage of how many rules are recovered with the overall sensitive rules. Ghost Rule Generator determines the percentage of how many rules are additionally generated with overall non sensitive rules. Given MST=30% and MCT=80% based on that WEKA tool gives the association rules. A transactional dataset made up of nine transactions with set of items. Given a transactional dataset D, Apriori algorithm applied on the dataset D, so that generates the association rules. Calculate the support and confidence for the transactional dataset. Based on the support threshold and minimum confidence threshold identify the sensitive rules. Apply the algorithms DSRRC, MDSRRC, MAXWT1 and MAXWT2 algorithms for hiding the data .During the hiding process calculate the values of the parameters. Table 3 calculates the support values of the transactional dataset. Table 4 calculates the support and confidence values of association rules. Table 5 compares the parameter values of association rule hiding among the DSRRC,MDSRRC and MAXWT1 approaches. Tables 6 compares the hiding parameter values among the DSRRC and MDSRRC and MAXWT2 approaches.

Table 2. A Transactional dataset

Transaction ID	Items
T1	ABCDE
T2	BCD
T3	ACE
T4	ADE
T5	AC
T6	AD
T7	BCE
T8	ABC
T9	BCE

Table 3. A support values of transactional dataset

Item set	Support
A	6/9=66.66
B	5/9=55.55
C	7/9=77.77
D	4/9=44.44
AB	2/9=22.22
AC	4/9=44.44
AD	3/9=33.33
BC	5/9=55.55
CE	4/9=44.44
DE	2/9=22.22
ABE	1/9=11.11
ACE	2/9=22.22
BCD	2/9=22.22
ADE	2/9=22.22
BCE	3/9=33.33
ABC	2/9=22.22
ABCD	1/9=11.11
ABCDE	1/9=11.11

Table 4. A support and confidence values of association rules

Association Rule	Support	Confidence
A->B	2/9=22.22	2/6=33.33
B->C	5/9=55.55	2/5=40
A->BC	2/9=22.22	2/6=33.33
AB->C	2/9=22.22	2/2=100
BC->D	2/9=22.22	2/5=40
BC->E	3/9=33.33	3/5=60
A->D	3/9=33.33	3/6=50
C->E	4/9=44.44	4/7=57.14
D->E	2/9=22.22	2/4=50

Table 5. Comparison Table of MAXWT1

Parameter	DSRRC	MDSRRC	MAXWT1
Hiding Failures(HF)	0%	0%	0%
Misses cost(MC)	38%	25%	15%
Artificial patterns(AP)	0%	0%	0%
Dissimilarity(DISS)	7.4%	6.4%	5.4%

Table 6. Comparison Table of MAXWT2

Parameter	DSRRC	MDSRRC	MAXWT2
Hiding Failures(HF)	0%	0%	0%
Misses cost(MC)	38%	25%	12%
Artificial patterns(AP)	0%	0%	0%
Dissimilarity(DISS)	7.4%	6.4%	5.1%
Side Effect Factors(SEF)	38.5%	27%	20%
Lost Rule Recovery(LRR)	70%	80%	88%
Ghost Rule Generation(GRG)	30%	15%	3%

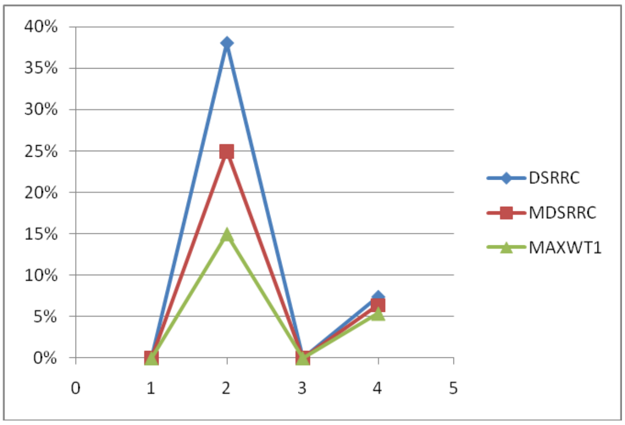

Fig.3. Comparison among DSRRC,MDSRRC and MAXWT1.

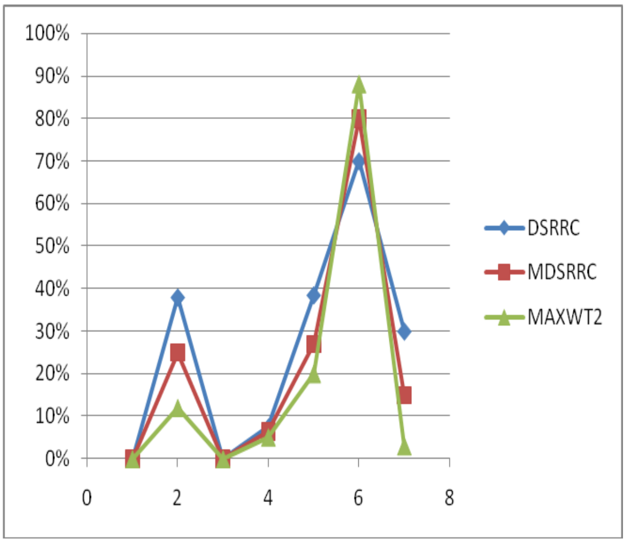

Fig.4. Comparison among DSRRC,MDSRRC and MAXWT2.

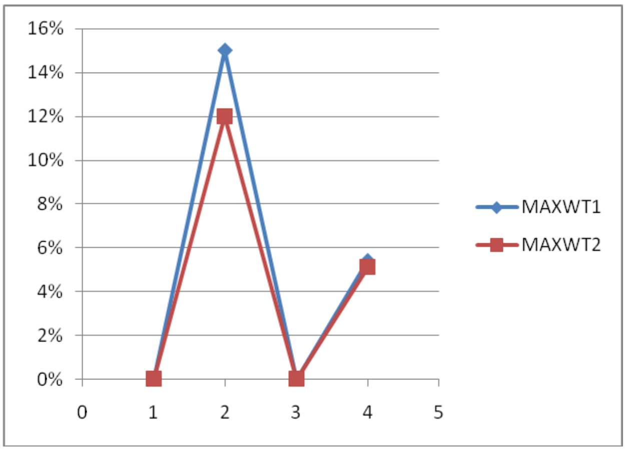

Fig.5. MAXWT1 vs MAXWT2.

VII. Conclusion

In this paper proposed novel algorithms that inherits the properties of CPT algorithm that gives better performance in hiding the rules. This algorithm works with binary transactional dataset and divides into blocks for generation of sanitized dataset. Block technique gives better performance in hiding rules. Additional parameters are also incorporated in MAXWT1 and MAXWT2 Algorithm. In future research work towards the direction of finding the more parameters and complexities of these algorithms.

References

[1] Rakesh Agrawal, Tomasz Imielinski, Arun Swami," Mining association rules between sets of items in large databases". ACM SIGMOD international conference on Management of data SIGMOD, pp. 207 (1993).

[2] Michael Hahsler,Bettina Grun,Kurt Hornik, "arules- A Computational environment for mining association rules and frequent itemsets" . Journal of Statistical Software. (2005).

[3] Pang-Ning Tan,Michael Steinbach,Vipin Kumar " Chapter 6. Association Analysis: Basic Concepts and Algorithms" . Introduction to Data Mining.

[4] Goldberg, David (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley Professional. ISBN 978-0201157673.

[5] Goldberg, David (2002). The Design of Innovation: Lessons from and for Competent Genetic Algorithms. Norwell, MA: Kluwer Academic Publishers. ISBN 978-1402070983.

[6] M.Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, "Disclosure limitation of sensitive rules".45–52.(1999).

[7] Elena Dasseni, Vassilios S. Verkios,Ahmed K. Elmagarmid,Elisa Bernito" Hiding association rules by using confidence and support".369–383.(2000).

[8] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni," Association rule hiding".434–447,(2004).

[9] Stanley R. M. Oliveira, Osmar R. Zaiane," Privacy Preserving Frequent Itemset Mining".(2002).

[10] Komal Shah, Amit Thakkar,Amit Ganatra, "Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S". (2012).

[11] Damandiya,UdayPratap Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database",pp-1306-1310.(2013).

[12] Tzung-Pei Hong,Chun-Wei Lin,Kuo-Tung Yang,Shyue-Liang Wang," A Heuristic Data-Sanitization Approach Based on SIF-IDF".(2011).

[13] Chun-Wei Lin, Tzung-Pei Hong, and Hung-Chuan Hsu," Reducing Side Effects of Hiding Sensitive Item sets in Privacy Preserving Data Mining".(2014).

[14] Narges Jamshidian, Ghalehsefidi,Mohammad Naderi Dehkordi, "A Hybrid Algorithm based on Heuristic Method to Preserve Privacy in Association Rule Mining".(2016).

[15] Belwal R., Varshney J, Khan S," Hiding sensitive association rules efficiently by introducing new variable hiding counter". (2013).

[16] C. N. Modi, U. P. Rao, and D. R. Patel," Maintaining privacy and data quality in privacy preserving association rule mining".(2010).

[17] C. N. Modi, U. P. Rao, and D. R. Patel," An Efficient Solution for Privacy Preserving Association Rule Mining", pp- 79–85, (2010).

[18] B.Kesava Murthy, Asad M.Khan, "Privacy preserving association rule mining over distributed databases using genetic algorithm" pp-S351–S364. (2013).

[19] Chun-Wei, Jerry Lin," A sanitization approach for hiding sensitive Itemsets based on particle swarm optimization", (2016).

[20] Mahtab Hossein Afshari," Association rule hiding using cuckoo optimization algorithm", 340–351. (2016).

[21] T.Satyanarayana Murthy, "Pine Apple Expert System Using Improved C4.5 Algorithm", pp-1264-1266,(2013).

[22] T.Satyanarayana Murthy, "Privacy Preserving for expertise data using K-anonymity technique to advise the farmers", (2013)

[23] N.P.Gopalan, T.Satyanarayana Murthy, "Association Rule Hiding Using Chemical reaction Optimization".(2017)

[24] N.P.Gopalan, T.Satyanarayana Murthy, Yalla Venkateswarlu,"Hiding Critical Transaction using Unrealization approach".(2017)

[25] T.SatyanarayanaMurthy, N.P.Gopalan, Yalla Venkateswarlu,"An Efficient Method for Hiding Association Rules with Additional Paramter metrics" .(2017)

[26] Yu-Chee Tseng and Hsiang-Kuang pan,"Secure and Invisible Data Hiding in 2-color Image".

Authors' Profiles

N.P.Gopalan: Professor at Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India. He obtained his Ph.D. from Indian Institute of Science, Bangalore, India. His research interests are in Data Mining, Distributed Computing, Cellular Automata, Theoretical Computer Science, Image Processing and Machine Intelligence.

T.Satyanarayana Murthy: Pursuing Ph.D. at Department of Computer Applications, National Institute of Technology, Tiruchirappalli. He obtained his B. Tech. in Information Technology from Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, India in 2006 and M. Tech. in Computer Science & Engineering from Andhra University, Visakhapatnam, Andhra Pradesh, India in 2010. His current research interests include Data Mining, Privacy Preserving, CBIR, Big Data, Soft Ccomputing, Optimization Techniques and Iintelligent systems. He is a member of SCRS society.

How to cite this paper: T.Satyanarayana Murthy, N.P.Gopalan," A Novel Algorithm for Association Rule Hiding", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.10, No.3, pp. 45-50, 2018. DOI: 10.5815/ijieeb.2018.03.06