A Survey On Various Techniques of Association Rules Hiding in Privacy Preservation Data Mining

Chandrima R Ghosh[1], Madhushree B[2]
[1, 2] Department of Computer Engineering
[1, 2] L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India

Abstract- *Data mining is the useful technology to extract information or knowledge from large database. However, misuse of this technology may lead to the disclosure of sensitive information. Privacy preserving data mining (PPDM) is new research direction for disclosure of sensitive knowledge. There are many techniques used in PPDM to hide association rules and generated by association rule generation algorithms. Association rule hiding is the method of modifying original database to make the sensitive rules disappear. The Security and Privacy of the data are main challenging issues. The owner of the data has some private information like association rule contained old database. However the integrity of mining results can affects gravely in case of unreliable administration supplier.*

Keywords- Association Rule Hiding,Privacy Preservation Data Mining ,Data Mining

I. INTRODUCTION

Data Mining is the process of extracting useful knowledge from large amounts of data. Data should be manipulated in such a sensitive way that information cannot be found through Data Mining techniques .While handling sensitive information it becomes very important to protect data against unauthorized access. This has increased the disclosure risks when the data is released to outside parties. This scenario leads to the research of sensitive knowledge hiding in database. [1]

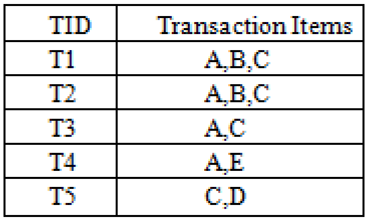
1.1 Privacy Preserving Data Mining (PPDM)

It is considered to maintain the privacy of data and knowledge extracted from data mining. It allows the protection of sensitive data or information while extracting. To preserve data privacy in terms of knowledge, one may modify the original database in such a way that the sensitive knowledge is not involved the mining result and non sensitive knowledge will be extracted. In order to protect the sensitive association rules, privacy preserving data mining include the area called "association rule hiding" [1]

1.2 Association Rule Mining

Association rule mining is the most effective data mining technique to discover hidden pattern from large volume of data. It was first introduced by R. Agarwal [12] in 1993. It works as follows: Suppose I = {i1, i2, ... , im } as a set of items, D = {t1, t2 , ... , tn} be a set of transactions where ti $\subseteq$ I. A unique identifier, TID, is associated with each transaction. A transaction t supports X, where a one set of items named as I, if X $\subseteq$ t. For example, let take a sample database of transactions.[1]

Table 1.Sample Transaction Table [1]



TID	Transaction Items
T1	A,B,C
T2	A,B,C
T3	A,C
T4	A,E
T5	C,D

An association rule is in the form X => Y, where X and Y are the subsets of item set in I, $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. In the rule X => Y, where X is called the antecedent (left-hand-side) and Y is the consequent (right-hand-side). Association rule mining generates high number of rules and only few of them are of interest. To solve the interested measurement problem, minimum support and minimum confidence thresholds are applied to each rule: Support for a rule X => Y, is denoted by S(X=>Y), is the proportion of transaction in the data set which contain the item set and is defined as[1]:

$$\text{Support}(X \Rightarrow Y) = |X \cap Y| \,/\, |D|,$$

Where $|X \cap Y|$ is the number of transaction containing the item set X and Y in the database, |D| denotes the number of the transactions in the data.

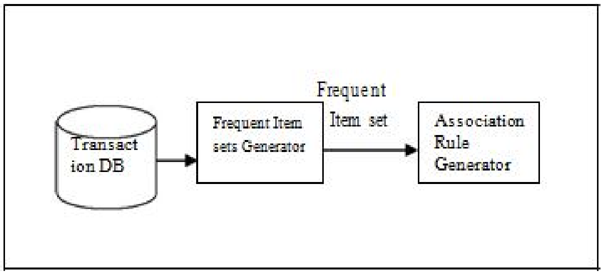
Confidence for a rule X => Y, is denoted by C (X =>Y), is taken as ratio of the support count of X union Y to that of the antecedent X defined as[1] :

$$\text{Confidence}(X \Rightarrow Y) = |X \cap Y| \,/\, |X|,$$

Where |X| denotes the number of the transactions in the database D that contains item set X. In other words, support describes how often the rule would appear in the database, while confidence measures the strength of the rule. A rule X=>Y is strong if support(X=>Y) ≥ minimum support and Confidence (X=>Y) ≥ minimum confidence

i. First find all frequent item sets- item set which occur at least as frequently as pre-determined minimum support count.[1]

ii. Generate stronger association rules which are based on the user defined minimum support and minimum confidence.[1]



There are different types of association rule mining algorithms which are available like Apriori algorithm, Partition algorithm, Dynamic item set counting algorithm, FP tree growth algorithm, etc. Apriori algorithm is one of the most popular and best-known algorithm to mine association rule, proposed by Agrawal and Srikant [1]. It makes user of prior knowledge of frequent itemset properties, which is a two-step process: join step and prune step. It moves upward in the lattice starting from level1 till level k and in reult there is no candidate set will remains after pruning.

1.3 Association Rule Hiding for PPDM

Association Rule hiding is the process of hiding strong association rules and creating sanitized database from the original database in order to prevent unauthorized party to generating frequent sensitive patterns. The problem can stated as " Given a transactional database D , minimum confidence as well as minimum support and a set R of rules can be mined from database D." A subset of R is denoted as set of sensitive association rules which Are to be hidden. The objective is to transform D into a database D' in such a way that no association rule in subset will be mined and all non sensitive rules in R could still be mined from D'

A rule for example X => Y, can be done by two ways it can be either by decreasing the support of the item set X and Y can below the minimum support threshold or decreasing the confidence of the item set X and Y can below minimum confidence threshold. Decreasing the confidence of a rule X => Y can be done by either increasing the support of X in transactions and not of Y or by decreasing the support of Y in transactions supporting both XY. Decreasing the support of a rule X => Y can be done by decreasing the Support of the corresponding large item set XY. [1]

Association rule hiding must satisfy some conditions which are given below:

i. Sensitive rule should not be generated from database.[1]

ii. Non sensitive rule must be generated from Sanitized database.[1]

iii. No new rule which is present in database should be generated from Sanitized database.[1]

1.4. Association Rule Hiding Approaches

Association Rule Hiding approaches can be classified into five classes which is discussed below: [1]

1.4.1 Heuristic Based Approaches is further divided into two techniques: i) Data distortion technique and ii) Data Blocking Technique.

A. Data distortion technique:- In this technique we replace 1- values to 0-values (delete items) or 0-values to 1- values(add items). There two approaches for rule hiding in data distortion based technique. First is reducing the confidence of rules and second is reducing the support of rules.

Table 2. Hiding A=>D by Distortion[1]

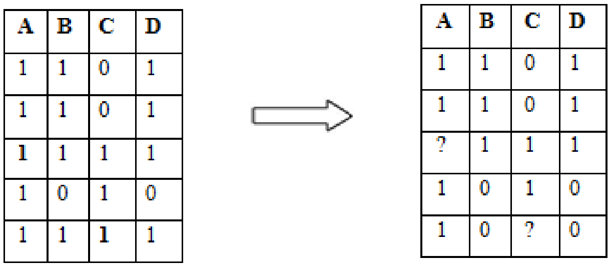
A	B	C	D
1	1	0	1
1	1	1	1
0	0	1	**0**
0	1	1	0
1	1	0	1

⇨

A	B	C	D
1	1	0	1
1	1	1	1
0	0	1	**1**
0	1	1	0
0	1	0	1

B. Data Blocking Technique is used to increase or decrease the support of the items by replacing 0's or 1's by unknowns "?", so that it become difficult for an adversary to know the value behind "?". This technique is effective and provides certain privacy. When hiding many of the rules at one time then they require less number of database scans and prune more number of rules.

Table 3. Hiding A=>C by Blocking[1]



A	B	C	D
1	1	0	1
1	1	0	1
1	1	1	1
1	0	1	0
1	1	1	1

A	B	C	D
1	1	0	1
1	1	0	1
?	1	1	1
1	0	1	0
1	0	?	0

1.4.2 Border Based Approach

Its hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent item sets of the original database. The item sets which are at the position of the borderline separating the frequent and infrequent item sets forms the borders. It uses the border of non-sensitive frequent item and computes the positive and negative borders in the item set. [1]

1.4.3 Exact Approach

It contains non heuristic algorithms which formulates the hiding process as a constraints satisfaction problem or an optimization problem which is solved by integer programming .In this approach minimally extends the original database by a synthetically generated database called extended database and formulates the construction of the extended database as a constraint satisfaction problem (CSP) which is then solved by using Binary Integer Programming (BIP) and the solution for association rule hiding is nothing but determining a sanitized database by satisfying constraints. [1]

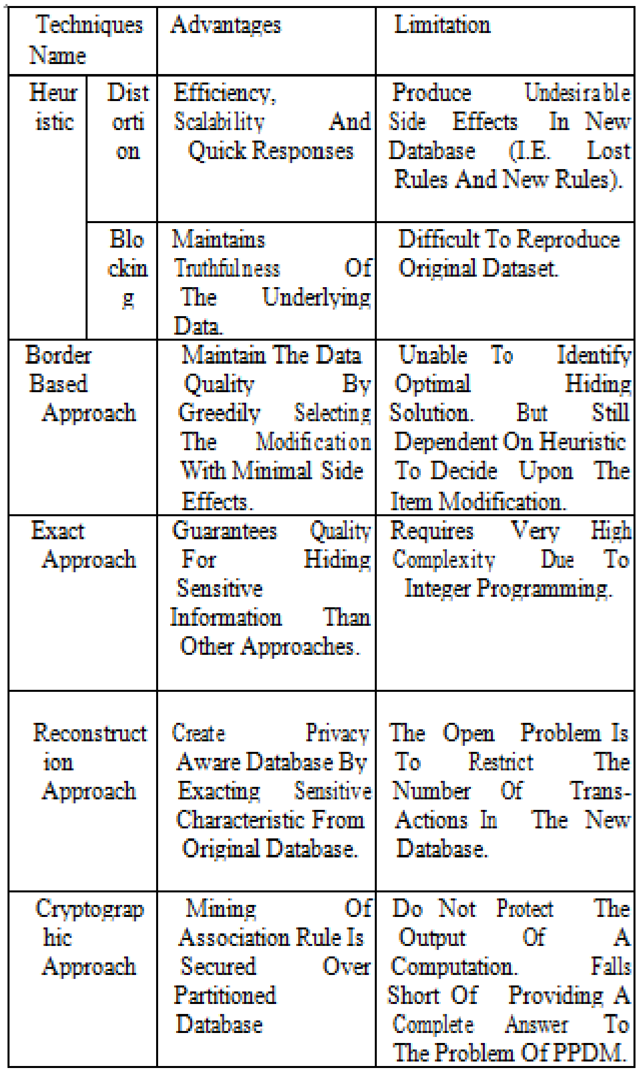
1.4.4 Reconstruction Based Approach

It is implemented by perturbing the data first and reconstructing the distributions at an aggregate level in order to perform the association rules mining. In which first places the original data aside and start from knowledge base. This approach has three phases in which first phase can generates frequent item sets from the original database and second phase performs sanitization algorithm over frequent item sets by selecting hiding strategy and identifying sensitive frequent items sets according to sensitive association rules. The third phase generates sanitized database by using inverse frequent item set mining algorithm and then releases this database. [1]

1.4.5 Cryptography Based Approach

It is used for multiparty computation, when database is distributed among several sites. Multiple number parties may share their private data without leaking any sensitive information at their end. It is divided into two categories: vertically partitioned distributed data and horizontally partitioned distributed data. In these approaches instead of distorting the database, it encrypts original database itself for sharing. The communication cost of this approach is very effective. [1]

Table 4. Comparative Study of existing approaches [4]



Techniques Name		Advantages	Limitation
Heuristic	Distortion	Efficiency, Scalability And Quick Responses	Produce Undesirable Side Effects In New Database (I.E. Lost Rules And New Rules).
	Blocking	Maintains Truthfulness Of The Underlying Data.	Difficult To Reproduce Original Dataset.
Border Based Approach		Maintain The Data Quality By Greedily Selecting The Modification With Minimal Side Effects.	Unable To Identify Optimal Hiding Solution. But Still Dependent On Heuristic To Decide Upon The Item Modification.
Exact Approach		Guarantees Quality For Hiding Sensitive Information Than Other Approaches.	Requires Very High Complexity Due To Integer Programming.
Reconstruction Approach		Create Privacy Aware Database By Exacting Sensitive Characteristic From Original Database.	The Open Problem Is To Restrict The Number Of Trans-Actions In The New Database.
Cryptographic Approach		Mining Of Association Rule Is Secured Over Partitioned Database	Do Not Protect The Output Of A Computation. Falls Short Of Providing A Complete Answer To The Problem Of PPDM.

II. RELATED WORK

Domadiya (2013)[4] described a heuristic based algorithm for hiding sensitive association rule, the algorithm named as Modified Decrease Support of R.H.S. item of Rule Clusters (MDSRRC) which is basically the algorithm is the modification of algorithm DSRRC and overcome the limitation of DSRRC, it is able to hide the sensitive association rule that contain multiple items in right hand side. The main advantage of proposed algorithm is, it does not make major changes in the database and it also able to hide rule which contain multiples item in right hand side of the rule.

Jadav (2013)**[5]** describes that database containing sensitive knowledge must be protected against unauthorized access. It has become necessary to hide sensitive knowledge in database. To address this problem, Privacy Preservation Data Mining (PPDM) include association rule hiding method to protect privacy of sensitive data against association rule mining. Various existing approaches to association rule hiding have been surveyed along with some open challenges.

Modi (2013)[2] proposed a heuristic algorithm named Decrease Support of R.H.S. item of Rule Clusters in short DSRRC ,which was able to hide many sensitive association rule at a time. They have analyzed experimental results for DSRRC, which show that performance of the DSRRC algorithm is better than other existing heuristic approaches. They have achieved improvement in misses cost, art factual patterns, dissimilarity and maintain data quality. This approach was able to hide only those rules that come on the right hand side (R.H.S.) contain single item, of the rule.

Weng[3]proposed an efficient algorithm, FHSAR which is used for fast hiding sensitive association rules. The algorithm can completely hide any given SAR by scanning database only one time and which significantly reduces the time while execution. Experimental results also show that FHSAR performs better than previous works in terms of execution time required and side effects which are generated in most cases.

Pathak (2012)[11] proposed an approach that is based on concept of pc cluster which improve performance by running operations in parallel, impact factor of a transaction which is equal to number of item sets that are present in those item sets which represents sensitive association rule and hybrid algorithm which is a combination of ISL (Increase support of LHS) and DSR (Decrease support of RHS). This approach is able to reduce the execution time and maintain data quality.

Sharma et al.(2014)[8] proposed an approach that is based on the combined techniques of randomization and k-anonymization. is divded into two algorithms. In algorithm I randomization is performed on dataset using attribute transitional probability matrix and in algorithm II k-anonymity is performed on randomized dataset which is result of algorithm I.

III. EXISTING ALGORITHMS

3.1 ISL (Increase support of LHS) and DSR (Decrease Support of RHS) [11]

In ISL method, Confidence of a rule is decreased by increasing the support value of Left Hand Side (L.H.S.) of the rule. For this, the items which are only from L.H.S. of a rule are chosen for modification. Whereas in DSR method, confidence of a rule is decreased by decreasing the support value of Right Hand Side (R.H.S.) of a rule. And for this purpose, only those items are chosen which are from R.H.S. of a rule for modification. But in both of the approaches required only less number of databases scanning and prune more number of hidden rules. More number of rules can be finding by database scanning.

3.2 Association Rule Hiding using Hidden Counter [1]

A heuristic based algorithm for addressing the problem of association rule hiding .they proposed a method for hiding sensitive association rule based on ISL (Increase the support of the item which is in the left hand side of the rule).they modified the definition of support and confidence, they introduced the use of a hidden counter in determining support and confidence.

3.3 FHSAR (Fast Hiding Sensitive Association Rules)[3]

An algorithm for hiding sensitive association rules is named as FHSAR (Fast Hiding Sensitive Association Rules) used for fast hiding sensitive association rules. It is able to hide any of the given sensitive association rule completely by scanning database only one time , which significantly reduces the time of execution. In this algorithm correlations between the sensitive association rules and each transaction are analyzed in the given original database , which can effectively select the proper item which is to be modify .

3.4 DSSR(Decrease support of R.H.S items in rule clusters) [2]

In this algorithm using given minimum support threshold (MST) and minimum confidence threshold (MCT), first generates the possible number of association rules from source database D. After that some of the generated association rules are selected as sensitive rule set by database owner. Rules with only single R.H.S. item are specified as sensitive. Then it finds C clusters based on common R.H.S. item in sensitive rule set and calculates the sensitivity of each cluster. After that it will index sensitive transactions for each cluster and sorts all of them by decreasing order of their sensitivities. From them to get the highest sensitive cluster, algorithm sorts sensitive transaction in the decreasing order of their sensitivities.

3.5 MDSRRC (Modified Decrease Support of R.H.S item of Rule Clusters) [9]

In this algorithm to hide the sensitive rule like X→Y, we can decrease either confidence or support of the rule below the user specified minimum threshold. This algorithm hides rules with multiple items in L.H.S and multiple items in R.H.S. So the rule is like aX→bY where a,bℵI and X,Y taht si b ereH .I item which is selected by this algorithm to decrease the support of the R.H.S. and confidence of the rule below minimum confidence threshold (MCT). We replace ‘1’ to ‘0’ in some transaction to decrease the support of selected items.

3.6 Random Perturbation [7]

In random perturbation the privacy of data can be protected by perturbing the sensitive data with randomization algorithm before releasing it to the data miner. The original data is distorted through adding the noise component to the data which is obtained through randomization. This method deals with character type, classification type, boolean type and number type of discrete data. To facilitate the conversion of data sets, it is necessary to preprocess the original data set. The method does not reconstruct the original data values, and it's only reconstruct distribution.

3.7 Method of k – anonymity [7]

When micro data is released for the research purpose, one needs to limit disclosure risk while maximize the utility of data. It introduced the k-anonymity technique to limit the disclosure risk . K - anonymity requirements says that, a data set is k anonymous (k 2: I) if each record in the data set is indistinguishable from at least (k-l) other records within the same data set. This k-anonymity requirement is generally achieved by using generalization and suppression. In generalization the attribute values are generalized in a particular interval. In suppression the attribute values are replaced or modified with some other values. Suppression contains information loss so it is generally avoided.

3.8 K-Means algorithm [6]

The standard k means algorithm selects k objects randomly from population as the initial centroid. If different initial values are given for the centroid, the accuracy output by the standard k-means algorithm can be affected. Usually a better clustering will give a group of centroids in which each centroid represents a group of similar objects. Suppose that if the initial centroids in which each centroid represent a group of similar objects can be given, then a better clustering can be obtain. The aim of k-means clustering algorithm is to partition objects into several classes and to make the distances between objects in the same class closer than the distances between the objects in different classes. So if certain centroids in which each centroid represents a group of similar objects can be obtained, then the centroids consistent with the distribution of data can be obtained.

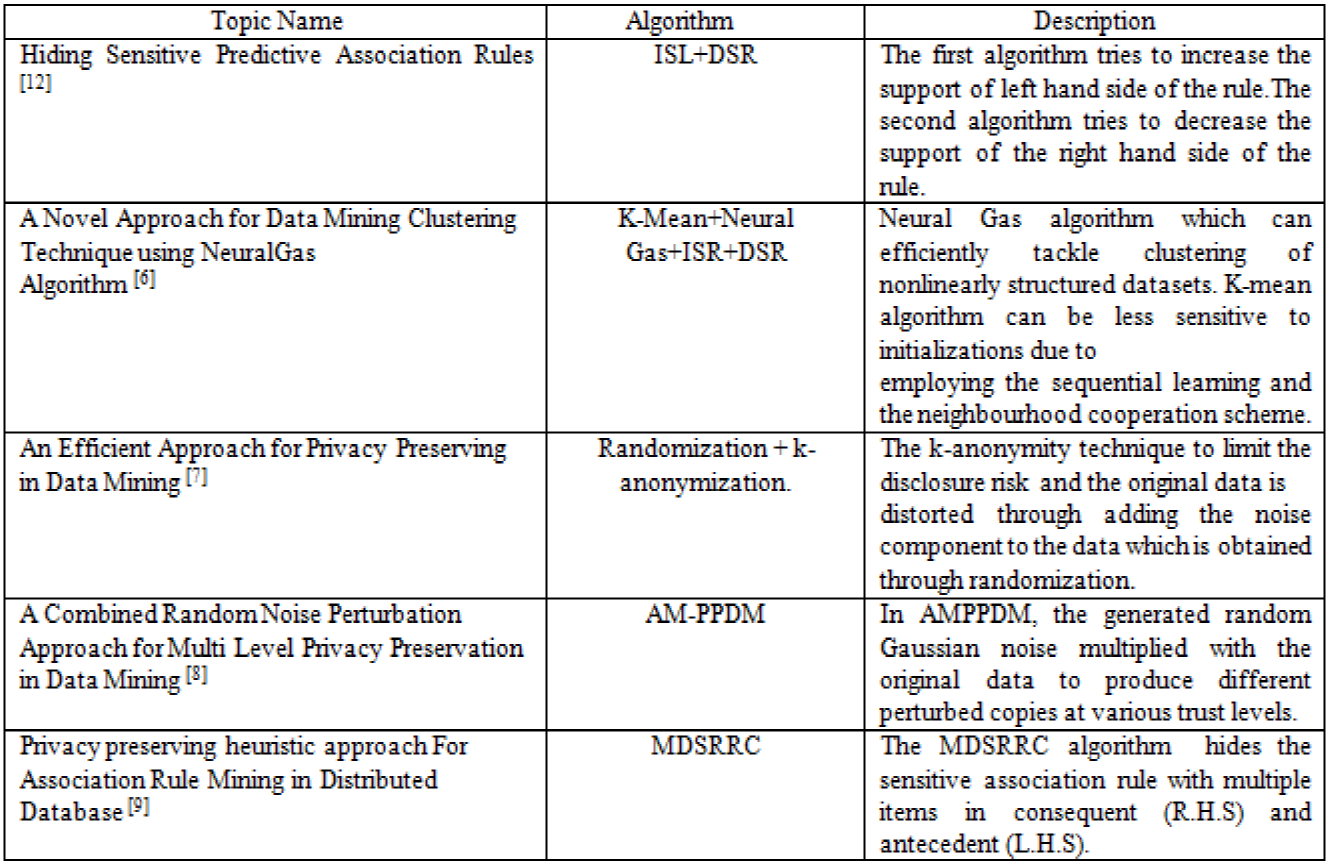
3.9 Neural Gas Clustering [6]

Neural gas is an artificial neural network, inspired by the self-organizing map (SOM). Neural gas is a simple algorithm for finding optimal data representations based on features. This algorithm was coined "neural gas" because of the dynamics of the feature vectors during the adaptation process which distribute themselves like a gas within the dataspace.

3.10 AM-PPDM(Additive Multiplicative Perturbation Privacy Preserving Data Mining) [8]

In this system, additive perturbation based PPDM is implementated in Multilevel trust The MLT-PPDM produces various perturbed set of the identification data for various levels. It prevents from diversity attacks, i.e., data miners can jointly reproduce the original data more accurately by comparing with the owner allowed data. Prevention of diversity attack can be done by appropriately making relationship with noise across at various trust levels. The noise covariance matrix has corner wave property. The data miners have no diversity gain during reconstruction.

IV. RESEARCH PAPER COMPARISON



Topic Name	Algorithm	Description
Hiding Sensitive Predictive Association Rules [12]	ISL+DSR	The first algorithm tries to increase the support of left hand side of the rule.The second algorithm tries to decrease the support of the right hand side of the rule.
A Novel Approach for Data Mining Clustering Technique using NeuralGas Algorithm [6]	K-Mean+Neural Gas+ISR+DSR	Neural Gas algorithm which can efficiently tackle clustering of nonlinearly structured datasets. K-mean algorithm can be less sensitive to initializations due to employing the sequential learning and the neighbourhood cooperation scheme.
An Efficient Approach for Privacy Preserving in Data Mining [7]	Randomization + k-anonymization.	The k-anonymity technique to limit the disclosure risk and the original data is distorted through adding the noise component to the data which is obtained through randomization.
A Combined Random Noise Perturbation Approach for Multi Level Privacy Preservation in Data Mining [8]	AM-PPDM	In AMPPDM, the generated random Gaussian noise multiplied with the original data to produce different perturbed copies at various trust levels.
Privacy preserving heuristic approach For Association Rule Mining in Distributed Database [9]	MDSRRC	The MDSRRC algorithm hides the sensitive association rule with multiple items in consequent (R.H.S) and antecedent (L.H.S).

V. CONCLUSION

I surveyed recent Association Rules Hiding Techniques and identified research trends. It showed that the K-mean and Neural-Gas used for hiding association rule having lowest time execution and AM-PPDM technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack.and MDSRRC modified version overcomes the limitation of DSSRC.

REFERENCES

[1] Kenampreet Kaur, Meenakshi Bansal"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.

[2] C. N. Modi, U. P. Rao, and D. R. Patel ., "Maintaining privacy and data quality in privacy preserving association rule mining," in Second International conference on Computing, Communication and Networking Technologies, pp. 1–6, Jul. 2010.

[3] C.Weng, S. Chen, H. Lo, "A Novel Algorithm for Completely Hiding Sensitive Association Rules," IEEE – 2008 Intelligent Systems Design and Applications, vol 3, pp.202-208.

[4] N. Domadiya and U.P. Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database," 3rd IEEE International Advance Computing Conference (IACC), pp. 1306-1310, 2013

[5] K.B.Jadav, J. Vania, D R. Patel, "A Survey on Association Rule Hi ding Methods," International Journal of Computer Applications (0975 – 8887), 82 (13), pp-20-25, 2013

[6] Mohnish Patel, Prashant Richhariya, Anurag Shrivastava "A Novel Approach for Data Mining Clustering Technique using NeuralGas Algorithm" IEEE,2014 pp. 251-254

[7] Manish Sharma, Atul Chaudhar,Manish Mathuria , Shalini Chaudhar, Santosh Kumar" An Efficient Approach for Privacy Preserving in Data Mining " IEEE,2014 pp.244-249

[8] Mr.S.Chidambaram , Dr.K.G Srinivasagan, “A Combined Random Noise Perturbation Approach for Multi Level Privacy Preservation in Data Mining” IEEE,2014 pp.1-5

[9] Bhoomika R Mistry, Amish Desai” Privacy preserving heuristic approach For Association Rule Mining in Distributed Database”IEEE.2015 pp.1-6

[10] K. Pathak, N. S. Chaudgari and A. Tiwari, “ Privacy Preserving Association Rule Miningby Introducing Concept of Impact Factor,” in 7th IEEE Conference on Industrial Electronics and Application(ICIEA), pp. 1458-1461, 2012.

[11] 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. 207-216,1993.

[12] Shyue-Liang Wang, Ayat Jafari “Hiding Sensitive Predictive Association Rules”IEEE-2005