Journal homepage www.jzs.uos.edu.krd
Journal of Zankoi Sulaimani
Part-A- (Pure and Applied Sciences)

Finding minimum confidence threshold to avoid derived rules in association rule mining

Nzar Abdulqader Ali

1 School of Administration and Economy, University of Sulaiman, New Campus, Sulaimaniyah, Kurdistan Region - Iraq
Email: nzar@mail.com

Article info
Original: 30 May 2015
Revised: 14 June 2015
Accepted: 25 June 2015
Published online:
20 Dec. 2015

Key Words:
Data Mining
Association Rule
Privacy Preserving

Abstract

Data in data warehouse often contains sensitive information, the concept of Privacy-Preserving has recently been proposed in response to the concerns of preserving sensitive information derived from published rules. A number of privacy preserving data publishing (PPDP) have been proposed. In this paper an algorithm proposed for hiding published rules that leads to disclosure of sensitive information by determining the confidence value of those rules from the raw data before running association rule mining using prior and posterior probabilities of generated rules and pass those confidence values to data miner to take it in his account when determining minimum confidence threshold in association rule mining algorithms .The experimental results show that the run time for deriving sensitive rules is stabile for different confidence values in comparison with other methods running linear programming methods for finding sensitive published rules. The most derived rules from goal rules (the rules derived from sensitive rules with minimum confidence value) located between 0.5 and 0.8 and these range of confidence values are critical values for data miner, finally experimental results shows that with support values %40,%58, and %63 still there is amount of derived published rules appears, and these results means that even with large minimum support threshold still derived published rules appears in association rule algorithms.

Introduction

Data mining is often represented as knowledge discovery in databases (KDD), and it is a process of nontrivial extraction of implicit, previously unknown and potentiality useful information from a large volume of data [1]. Every day huge amounts of data are being collected on individuals by government, organizations, and industries. Data mining techniques and its algorithms operates on the original data set and finally produce patterns which will cause the leakage of privacy data. At the same time the relationships between stored data tends to deriving sensitive information from published rules. These problems challenge the data miner to interest in preserving privacy when mining data [2].

Association Rule

The idea of association rules derived from the market basket analysis where a rule likes “When a customer buys a set of items what is the probability that he buys another item?” An association rule is a combination of Boolean conditions constructed from the attribute values of a dataset that occur together with greater frequency than might be expected if the conditions were independent of one-another. Mathematically, an

association rule is defined as $X \rightarrow Y$, where X (antecedent) and Y (consequent) are logical predicates constructed by Boolean predicates. In a transactional data set, an association rule appears as (*item = milk*) $\wedge$ (*item = sugar*) $\rightarrow$ (*item = bread*) which means when a customer buys milk and sugar, it is most likely that he also buys bread. Apriori is one of the most implemented algorithms of association rule mining. Apriori algorithm is controlled by two joint and conditional probabilities called support and confidence. The Support thresholds control the number of items represented in the rules and the confidence control the number of published rules [3].

Definition 1 (Support): It is the probability of item or item sets in the given transactional data base support(X) = n(X) / n where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set therefore.

$$\text{Support } (X \rightarrow Y) = \text{support } (XUY).$$

Definition 2 (Confidence): It is conditional probability, for an association rule $X \rightarrow Y$ and defined as

$$\text{Confidence}(X \rightarrow Y) = \text{support } (XUY) / \text{support}(X).$$

Definition 3 (Frequent itemset): Let A be a set of items, T be the transaction database and α be the user specified minimum support. An itemset X in A (i.e., X is a subset of A) is said to be a frequent itemset in T with respect to α, if $\text{support}(X)_T > \alpha$.

The problem of mining association rules can be decomposed into two sub-problems:

- Find all sets of items (itemsets) whose support is greater than the user-specified minimum support α such item sets are called frequent itemsets.
- Use the frequent itemsets to generate the desired rules. The general idea is that if, say ABCD and AB are frequent itemsets, then we can determine if the rule AB$\rightarrow$ CD holds by checking the following inequality

  $\text{support}(\{A,B,C,D\}) / \text{support}(\{A,B\}) \geq \epsilon$ where the rule holds with confidence ϵ [4]

Privacy Preserving Data Publishing

Information in data ware house represented in terms of identifiers. Privacy preserving data publishing (PPDP) starts with items included in different type of identifiers. These identifiers are in one of the following forms

- Identifiers: These are attributes that containing information that explicitly identifies record like social security number.
- Quasi-identifiers (QID): Represent a set of attributes used for linking with external information in order to uniquely identify individuals in a given table.
- Sensitive attributes (SA): These attributes contain values that are considered to be sensitive to the victim.
- Non-sensitive attributes (NSA) :Non-Sensitive attributes contains all attributes that do not fall into the previous three categories

The most important task of data miner in privacy field is distinguishing among these types of attributes especially the sensitive attribute(s) from the other types [5].

Related Works

There have been many approaches to privacy preserving data publishing (PPDP). In [6] the authors hide sensitive rules depending on the principle of maximum entropy function using optimization method, they prove that without optimization method the large amount of variables effect the non-linear programming software runs for about 30 seconds, before it reports “out of memory” error . The optimization can decrease

the number of variable dramatically and the solver software successfully finish. The drawback of this method is the time need for optimization method using maximum entropy function to prune items with minimum amount of information used by apriori algorithm, and this problem increase with increasing amount of items in the data warehouse. In [7], they proposed an association rule hiding algorithm for privacy preserving data mining. Their work based on association rule hiding by modifying the database transactions so that the confidence of the association rule can be reduces. Changing the value of confidence by transection modification tends to hide some rules that include non-sensitive information and may be important rules for publishing. In [8], the authors use distortion technique for hiding sensitive association rules. The algorithms that they used either hide a specific rule using data alteration technique or hide the rules depending on the sensitivity of the items to be hidden. The advantage of this algorithm doesn't modify the database transection so that the support and confidence of the association rules remain unchanged, In time removing the items from the rules make other rules hided from publishing by decreasing number of items in LHS or RHS of the rules.

Proposed Algorithm

A. Generating Association Rules

The regular algorithm for generating Association rules is failed in our case since the item sets in our study include more than one labels and the right hand side (RHS) is fixed for sensitive item(s) in previous studies to solve this problem they fixed the RHS by mixing the Association rule mining with classification method (ie. determining the sensitive attributes as class attributes) and the drawback of this algorithm appears in selecting only one item as a sensitive item since the class item fixed to only one item. In this paper an algorithm is proposed for generating rules depending only on the LHS items (non-sensitive items) and the remaining item(s) is/are represents the RHS (sensitive items) as shown in Algorithm 1

Algoriyhim 1: Generating Rules

```
Inputs:
DT: Dataset
SA: list of all sensitive attributes
Outputs:
PRL: All possible rules generated from the dataset
Begin
// find number of all tubles in DT
NOT = select count (*) from DT
// find all non-sensitive attributes
NSA= Select all attributes from sysobjects, syscolumns where the selected attributes not equal to  SA
For each items in NSA do
 n= Calculate total number of items in non-sensitive attributes
End
// combination of all items in the attributes (CAR)
   For r=1 to NSA
   CAR=CAR + Combination (n,r)
   End
// combination of all items in the same attribute (CSA)
  For each  NSA ϵ DT do
    For each INS ϵ NSA do
        For r=0 to INS
           CSA=CSA + Combination (INS,r)
        End
    End
 End
// combination of all possible rules generated from NSA
```

```
    POR=CAR-CSA
// generating all possible rules
For i= 1 to POR
 For each ITS ϵ NSA
  PRL=PRL + ITS
 End
End
```

B. Target Rules

After generating all possible rules depending on given attributes and item sets, we name the rules with confidence one (%100 published in generating rules) target rule, the idea behind selecting these rules as a target rules is that for any minimum confidence value determined by data miner these rules are published and there is no possibility for hiding them.

C. Goal Rule

The goal rule is one of the combination rules generated from target rule with Minimum confidence, the reason we select the goal rule with minimum confidence is to increase number of derived rules generated depending on goal rule. The steps for generating goal rule are represented in the algorithm-2

Algorithm-2: Generating Goal Rule

```
Inputs:
DT: Dataset
PRL: Published Rule
LAT: LHS of Attributes in PRL
RAT: RHS of Attributes in PRL
Outputs:
GRL:  Goal Rule
Begin
// find number of all tubles in DT
NOT = select count (*) from DT
// find combination of all rules with same LHS attributes of PRL
  For each  LAT ϵ  PRL  do
    For each INS ϵ LAT do
        For r=0 to INS
            CGRL= CGRL + Combination (INS,r)
         End
    End
 End
 // find frequency of all LHS  items in  CGRL in DT
LFCGRL= select count(*) of all items in DT  with same CGRL
// find frequency of all LHS and RHS items in  CGRL  in DT
LRFCGRL= select count(*) of all items in DT  with same LHS and RHS of CGRL
// Support of the Goal  rules
SUP= LFRCGRL / NOT
// Confidence of the derived rules
CONF=SUP/(LFCGRL/NOT)
// find goal rules among candidate goal rules (CGRL)
GRL=min confidence (CGRL)
End
```

D. Derived Rules

Derived rules are combination of all rules generated depending on goal rule. First the system find the rule that there sensitive attributes are complemented with the goal rules RHS sensitive attributes, and then the system depending on each item of the attributes in LHS find all combination of rules generated by knowing the LHS and RHS of selected items. The steps for generating derived rules are represented in the following algorithm.

Algorithim-3

```
Inputs:
DT: Dataset
PRL: All possible rules generated from the dataset
GRL: Goal Rule
LAT: LHS of Attributes in GRL
RAT: RHS of Attributes in GRL
Outputs:
DRL: All Derived rules generated from Goal Rule
Begin
// find number of all tubles in DT
NOT = select count (*) from DT
// find combination of all rules with same LHS attributes of goal rule but complement of RHS
GRLC=select LAT from PRL where RHS <> RAT
// find combination of all rules generated from LHS attributes of goal rule
  For each  LAT ϵ  GRL  do
    For each INS ϵ LAT do
        For r=0 to INS
           RGRL= RGRL + Combination (INS,r)
         End
    End
 End
// add goal rule complement to other combination of rules
DRL = GRLC+RGRL
// find frequency of all LHS  items in  DRL in DT
LFDRL= select count(*) of all items in DT  with same LAT
// find frequency of all LHS and RHS items in  GRLC in DT
LRFDRL= select count(*) of all items in DT  with same LAT and RAT
// Support of the Derived  rules
SUP= LFRDRL / NOT
// Confidence of the derived rules
CONF=SUP/(LFDRL/NOT)
End
```

To illustrate our proposed algorithm a prototype system is implemented using Microsoft SQL server database as backend data warehouse and visual basic.net as front end interface and to illustrate the concept of the proposed algorithm an example using real adult dataset from (UCI) UC Irvine machine learning repository data source [9] is used.

Example:
To illustrate the proposed algorithm from UCI data one of the generated rule with confidence one selected.

Work class= Federal-gov, Education= 10^{th} ⟶ salary <=50 conf.=1 ……..…..…(1)

In this case all possible rules generated are

Work class= Federal-gov ⟶ Salary <=50 conf.= 0.61 ……..... …… ..…… (2)
Education= 10^{th} ⟶ Salary <=50 conf.= 0.93 …… …….(3)

Since the confidence of rule (2) is less than the confidence of rule (3) we take the second rule as a goal rule. Rules (4) to (27) are the remaining rules derived from the Goal rule (2)

Work class= Federal-gov ⟶ Salary >50 conf.= 0.39……...….(4)
Work class= Federal-gov & education= 10^{th} ⟶ Salary <=50 conf.= 1 ……..….(5)
Work class= Federal-gov & education= 11^{th} ⟶ Salary >50 conf.= 0.11 ……..…(6)
Work class= Federal-gov & education= 11^{th} ⟶ Salary <=50 conf.= 0.89……...(7)
Work class= Federal-gov & education= 12^{th} ⟶ Salary <=50 conf.= 1 ……..….(8)
Work class= Federal-gov & education= 5^{th}-6th ⟶ Salary <=50 conf.= 1 ……..….(9)
Work class= Federal-gov & education= 9^{th} ⟶ Salary >50 conf.= 0.33 ……(10)
Work class= Federal-gov & education= 9th ⟶ Salary <=50 conf.= 0.67……...(11)
Work class= Federal-gov & education= association-acdm ⟶ Salary <=50 conf.= 0.65……...(12)
Work class= Federal-gov & education= association-acdm ⟶ Salary >50 conf.= 0.35 ……..…..(13)
Work class= Federal-gov & education= assoc-voc ⟶ Salary <=50 conf.=0.61……...(14)
Work class= Federal-gov & education= assoc-voc ⟶ Salary >50 conf.= 0.39 ……..(15)
Work class= Federal-gov & education= bachelors ⟶ Salary <=50 conf.= 0.55…….(16)
Work class= Federal-gov & education= bachelors ⟶ Salary >50 conf.= 0.45….……...(17)
Work class= Federal-gov & education= doctorate ⟶ Salary <=50 conf.= 0.07…,,,.…..(18)
Work class= Federal-gov & education= doctorate ⟶ Salary >50 conf.= 0.93…….…..(19)
Work class= Federal-gov & education= HS-grade ⟶ Salary >50 conf.= 0.28……..…..(20)
Work class= Federal-gov & education= HS-grade ⟶ Salary <=50 conf.= 0.72……..(21)
Work class= Federal-gov & education= maters ⟶ Salary <=50 conf.= 0.31……..(22)
Work class= Federal-gov & education= masters ⟶ Salary >50 conf.= 0.69……..….(23)
Work class= Federal-gov & education= prof-school ⟶ Salary <=50 conf.= 0.19……..(24)
Work class= Federal-gov & education= prof-school ⟶ Salary >50 conf.= 0.81……..…..(25)
Work class= Federal-gov & education= some-college ⟶ Salary >50 conf.= 0.33……..…..(26)
Work class= Federal-gov & education= some-college ⟶ Salary <=50 conf.= 0.67…….….(27)

The data miner depending on the confidence of derived rules decides the minimum confidence requested to hide the rule from publishing and as a result prevent privacy of the items. Figure-1 shows the implementation of our prototype system.

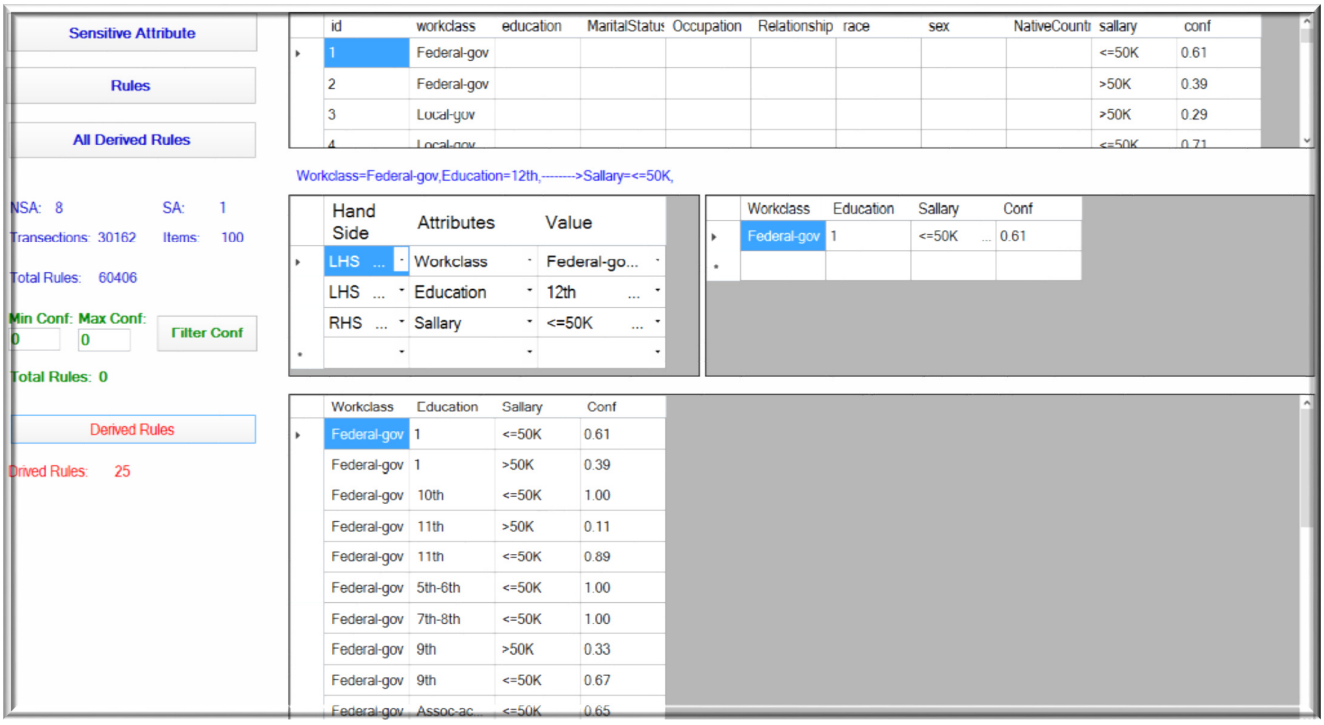

Figure-1: Prototype system implementation

Rustles and Discussions

One of the biggest challenges in data mining field is the algorithm execution time. In the proposed prototype system the average execution time needed for deriving privacy rules depending on different confidence values is calculated as shown in Figure-2.

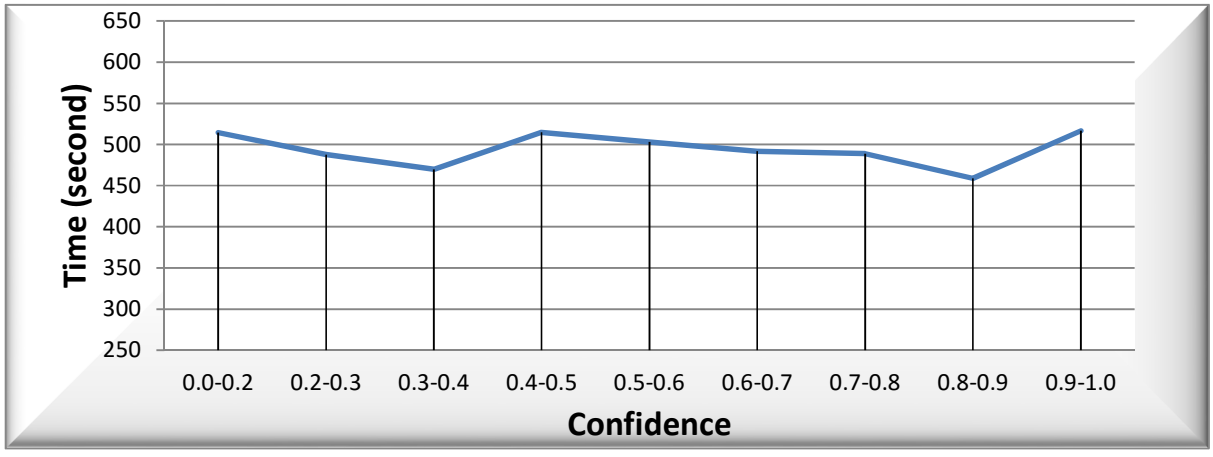

Figure-2: Average execution time

The average execution time is between 450 second to 520 seconds for different confidence values and approximately for 3000000 derived rules from 40719 published rules with confidence one. This result represents minimum average execution time in comparison with maximum entropy function as proposed by [6] and in the same time shows stability of the system with different confidence values.

Figure-3 represents average number of derived rules for different confidence values. It is clear that the confidence of the most derived rules from goal rules are between 0.5 and 0.8 and these range of confidence values are critical values for data miner as a minimum confidence threshold, which is means still we need hiding published rules that tends to derive sensitive information from them.

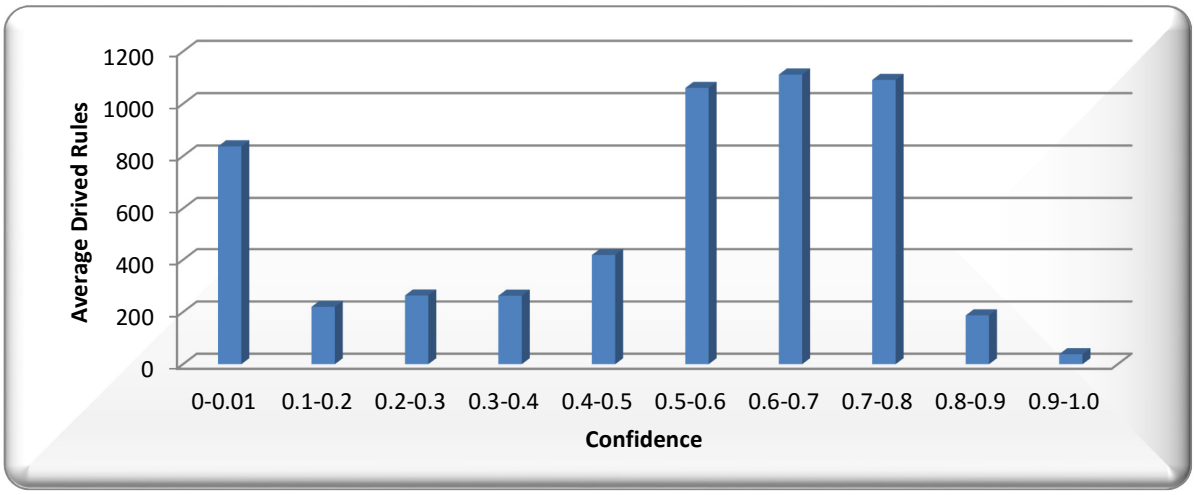

Figure 3: Average Derived rules

Figure-4 and Figure-5 shows that with support values 40% and 63%,still there is amount of derived published rules, and this conclusion means that even with large minimum support threshold values still there is possibility for derived published rules which violate individual privacy.

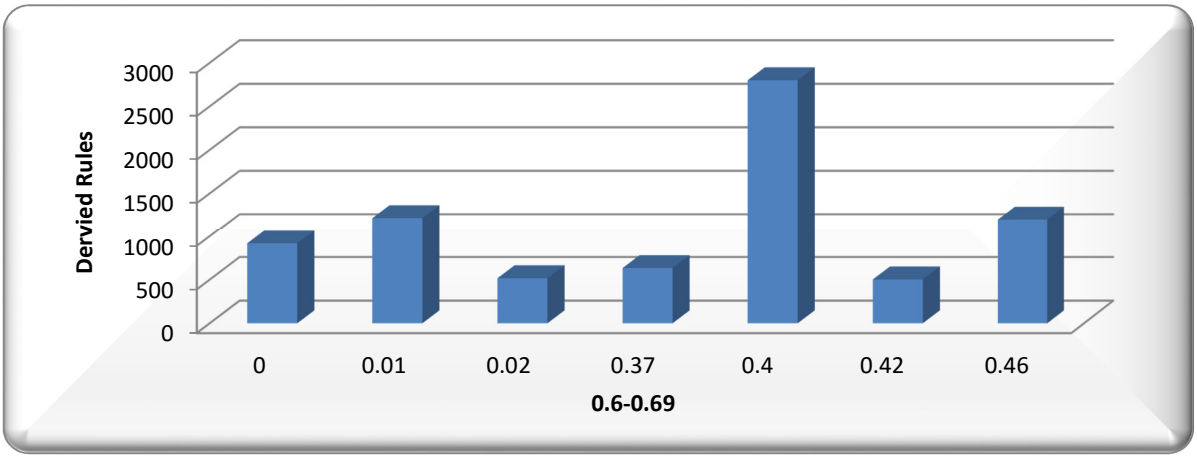

Figure-4 : Dervied Rules generated in support (0-0.46) with conf.(0.6-0.69)

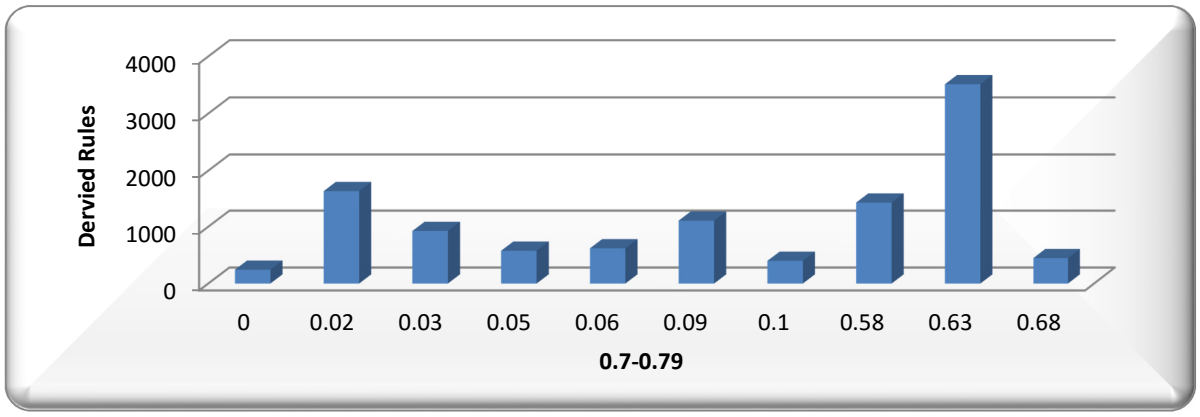

Figure-5 : Dervied Rules generated in support (0-0.68) with conf.(0.7-0.79)

Conclusion and Future works

An algorithm for hiding published rules that leads to disclosure of sensitive information by determining the confidence value of those rules from the raw data before running association rule mining is proposed. The experimental results shows the stability of execution time for difference confidence and with height minimum confidence and minimum support thresholds still there is need for hiding publish rules. Several directions of the future work can be followed. One direction is to scale the data warehouse to test the scalability, to measure the time cost for generating derived rules from Goal rules. Another interesting direction is to implement the same algorithm on different data types (transactional, relational, location, social, graphs) and each of these data types requires different techniques for pre-processing and mining rules.

References

[1] S. Sumathi and S.N. Sivanandam. "*Introduction to Data Mining Principles, Studies in Computational Intelligence*" (SCI) 29, 1–20 (2006).

[2] Andrei Manta. *"Literature Survey on Privacy Preserving Mechanisms for Data Publishing",* Msc, November 2013, URL: http: 1,2013, http:// cybercybersecurity.tudelft.nl/sites/default/files/ Literature_Survey_Andrei_Manta_0.pdf, accessed at: May,2015.

[3] Thi-Thiet Pham, Jiawei Lu, Tzung-Pei Hong, BayVo *"An Efficient Algorithm For Mining sequential Rules with Interestingness Measures"*, *International Journal of Innovative, Computing, Information and Control* ICIC International,Vol 9, Number **12**, pp 4812,December **2013**

[4] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, "Mining Frequent temsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010

[5] Deepa B. Mane ,Emmanuel M, "*Review on Privacy and Utility in High Dimensional Data Publishing*", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 1, January – February 2014.

[6] Zutao Zhu, Guan Wang, Wenliang Du. "*Deriving Private Information from Association Rule Mining Results"*. pp18, ISSN :1084-4627, E-ISBN : 978-0-7695-3545-6 IEEE,(2009).

[7] Sunil Kumar, Mahaveer Singh and Nidhi Porwal, "An Algorithm for Hiding Association Rules on Data Mining". National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC,(2012)

[8] K.Srinivasa Rao, CH. Suresh Babu, A. Damodaram and Tai-hoon Kim, "Distortion Technique for Hiding Sensitive Association Rules". International Journal of Multimedia and Ubiquitous Engineering Vol. 9, No. 10 (2014), pp. 57-66 (2014).

[9] UCI, Machine Learning repository, adult data set, available at: http://www.ics.uci.edu/mlearn/mlrepository.html , accessed at: May,2015.