International Journal of Control and Automation
Vol. 12, No. 4, (2019), pp. 29-36

Secure Multi Parties Communication using Randomize Group Technique in Privacy Preserving Data Mining

Mohd Ashraf

Associate Professor, CSE, School of Technology, Maulana Azad National Urdu University, Hyderabad Email: ashraf.saifee@gmail.com

Md. Zair Hussain

Associate Professor, Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad Email: mdzairhussain@gmail.com

Dinesh Kumar Singh

Assistant Professor, Department of Information Technology, Dr. Shakuntala Misra National Rehabilitation University Lucknow, UP, India, Email: dineshsingh025@gmail.com

Abstract: *The approach which used for extracting the patterns and rules from data is known as data mining. It is also called as Knowledge Discovery from Data or KDD process. Generally data mining approaches are applied on model of dataware house, in which data are collected into central sites and then run the algorithm. The security of private data is the major issue in recent years. Easy availability of personal data raises the issue of privacy preserving data mining. When data is transferred between third parties then it is very necessary to secure that data. Accuracy and privacy of the personal data is important concern in the field of current research. There are several techniques and methods are available for preserve the private data, but development in the fruitful direction provide the accurate and efficient data without any loss. In this way it is very important to develop the techniques in data mining which do the work without sacrifice its privacy concern. To deal with privacy issues of data, Privacy preserving data mining is using. There are many privacy preserving data mining approaches are available to convert the original data. The privacy preserving data mining approaches are examined through its privacy protection degree, accuracy, applicability and efficiency. The proposed research work using randomized group approach for preserving the data in privacy preserving data mining.*

Index Terms: *data mining; privacy preserving; technique; randomized group, secure.*

I. INTRODUCTION

In data mining process the private data preservation is very important. This is the research area where many researchers developed their research in several ways. To avoid the bias answer and to prevent the survey data it developed randomized response technique [1] [2]. Preserving the private data is very

ISSN: 2005-4297 IJCA

necessary in privacy preserving data mining.

II. Data mining fundamental

For knowledge discovery data managing capacities and improved data give new learning chances. In this decade inquire about on knowledge discovery in databases in Interdisciplinary way has risen. In the social insurance field the information of example acknowledgment should associated with dexterous individuals. Data mining is the procedure which has mechanized example acknowledgment. For finding the concealed examples the information mining have a few techniques that is connected to KDD and it is hard to find that examples with past quantifiable procedures. Based on how the example keeps the subtle cases the examples are determined. Information vaults, Databases and information stockrooms are getting the chance to be inescapable however it is required aptitudes and learning to get the advantages from this stored data [3].

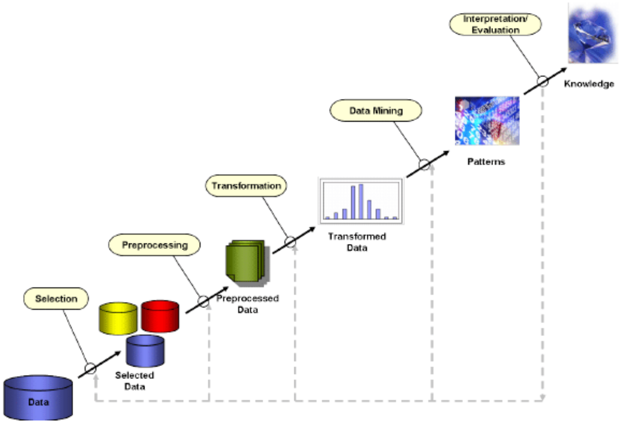


Figure 1: KDD process

III. Data mining Process

The steps of data mining process are discussed below:

1. Understanding of business

In the very first step it includes the requirement and objective of the business, and then to design the objective converts this knowledge into data mining problem.

2. Understanding of data

Fisrt of all it should collect the data and then to identify data quality problems, to detect interesting subsets then get familiar with that data.

3. Prepartion of data

To create the final data set all types of activities are combined called as data prepartion. This step include work realted to case, attribute selection, table, for modeling tool data clening etc.

4. Modelingin

This step several modelling techniques are select and apply ed. For same data mining problems it can apply different data modelling technique.

5. Evalution

The steps that use to evalue the steps and review the created model include in evaluation step, it is necessary to achieve a proper business target. If any business issue is not sufficinetly consider than determine that.

6. Deployment

In this step the result is organized and presented. It is as simple like generate a report or it can be difficult as developing a data mining process for reapetable data mining.

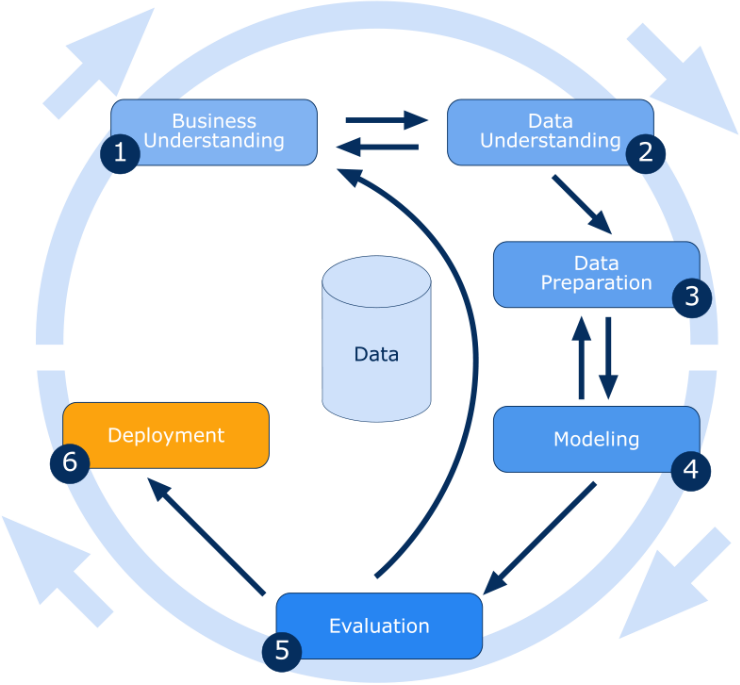


Figure 2: Steps of data mining process

IV. PRIVACY RELATED ISSUES IN DATA MINING

In numerous applications, it is think about that for gadget enabling divulgence of supportive examples data mining is earth shattering device. As there is huge information distribution center is accessible and it identified with various methodologies it is important to save the data in numerous conditions, for instance data about state of patient, singular establishment information and customer tendencies and so on. It unavoidably makes private information of the customer in the event that it reveals the private unique information. Along these lines the fundamental objective of data mining is to find the methodologies where information can get with protection safeguarding. Security protecting information digging use for tackle this issue change the first information there are numerous protection saving procedures are utilized. The security safeguarding information mining estimated based on measurements of security insurance, calculation, exactness and relevance [4].

In data mining huge database stored data related to sensitive factors. So it is require to prepare the data which can reveal that patterns and data which may bargain secrecy and protection commitments. The risk of disclosure the sensitive data is increased due to Progression of proficient data mining method. For this data aggregation is consider as a common way. Once data set is compiled the risk to a person's protection becomes an integral factor, cause the data miner. And it make it available for identify the specific people, when initially the information were mysterious initially.

V. PRESERVING THE PRIVACY IN DATA MINING

By increasing the sophistication of data mining algorithm and increasing the ability to store user's personal data in recent years in data mining privacy preserving is very important. In order to perform privacy preserving data mining many techniques like association rule mining, clustering, classification and k-anonymity are used in recent survey. For supporting variety of domains like medical diagnosis, national security, weather forecasting and retailing the data mining has developed successfully. For example in an healthcare center it is important to protect the personal data of patient [5]. As privacy concern enhancing the data mining becoming more pervasive.

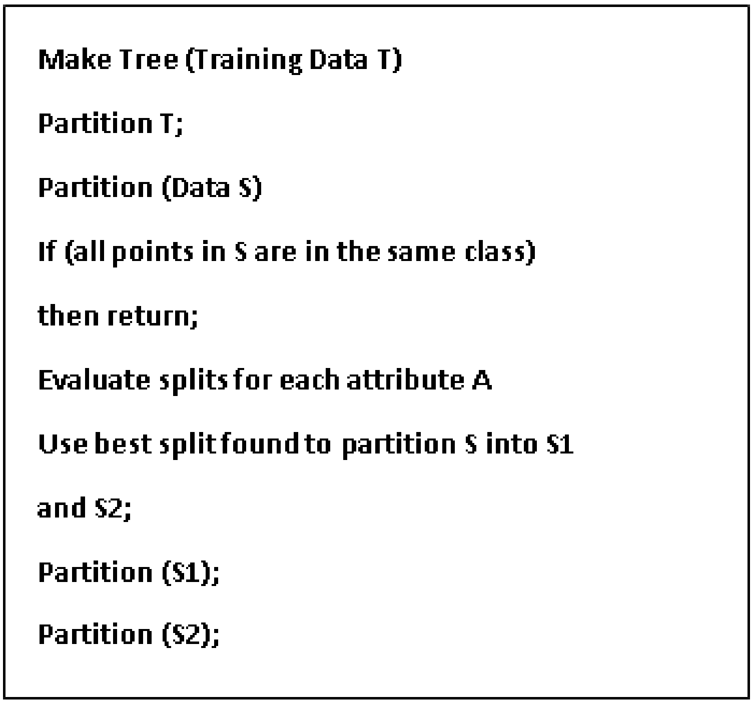
VI. DATA PREPROCESSING

For preprocessing of data decision tree is used for prediction algorithms. This approach uses ID3 and decision tree for classify the problems. With respect to variety of predictor types the decision tree algorithm is robust. To create the subset of possibly useful predictors the decision trees can use on the first pass of a data mining, and it is comparatively take less time. It can fed into normal statistical routines, nearest neighbor and nearest neighbor. If in the model there is large number of possible predictors then it can take considerable amount of time to run.

a. Decision Tree Classification

Tree building and tree pruning are the two classification of decision tree.

By repeatedly partitioning of training data the initial decisions tree is grown in tree building phase. By using an attribute the training set is divided in to two or more parts. Until the example of each part belongs to a class this process repeated.



```
Make Tree (Training Data T)
Partition T;
Partition (Data S)
If (all points in S are in the same class)
then return;
Evaluate splits for each attribute A
Use best split found to partition S into S1
and S2;
Partition (S1);
Partition (S2);
```

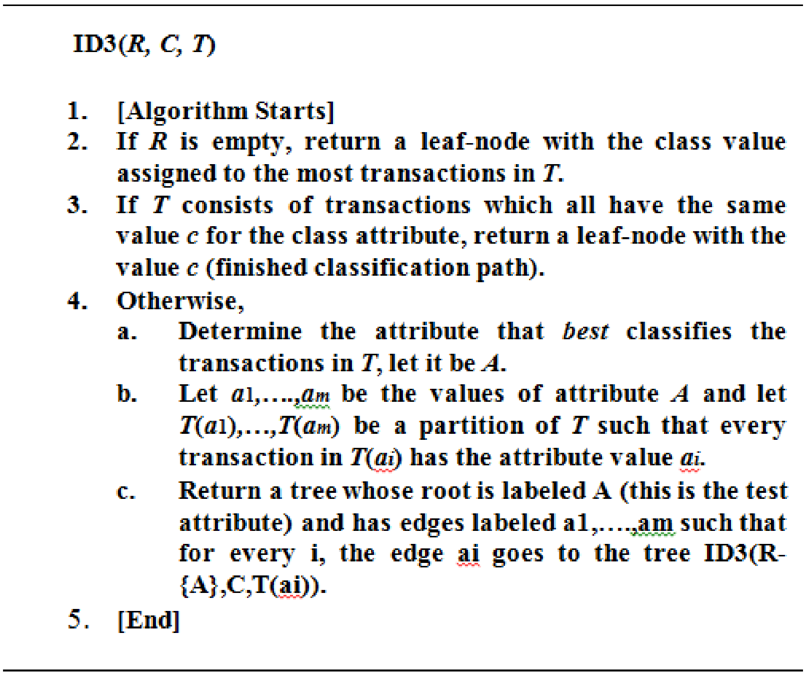
Figure 3: Algorithm for building a tree

b.ID3 Algorithm

In the concept of ID3 algorithm it is consider that each attributes contain discrete data with continuous data and divided in categories.

The idea of ID3 algorithm dependent on that each characteristic contain discrete information and it separated in classes, with portrayal of proceeds with information [6].

The ID3 algorithm is given beneath. In this algorithm tree is made in top down methodology in recursive way. Each property at the root watches that how trait can order the exchange. In these traits best qualities are picked and the remainder of the exchanges was partitioned by it. After that on each parcel ID3 is called recursively [7].



```
ID3(R, C, T)

1.  [Algorithm Starts]
2.  If R is empty, return a leaf-node with the class value
    assigned to the most transactions in T.
3.  If T consists of transactions which all have the same
    value c for the class attribute, return a leaf-node with the
    value c (finished classification path).
4.  Otherwise,
    a.  Determine the attribute that best classifies the
        transactions in T, let it be A.
    b.  Let a1,....,am be the values of attribute A and let
        T(a1),...,T(am) be a partition of T such that every
        transaction in T(ai) has the attribute value ai.
    c.  Return a tree whose root is labeled A (this is the test
        attribute) and has edges labeled a1,....,am such that
        for every i, the edge ai goes to the tree ID3(R-
        {A},C,T(ai)).
5.  [End]
```

Figure 4:: The ID3 Algorithm for Decision Tree Learning [8]

VII. Randomized Response Techniques

To secure the information of study examines built up a system called Randomized Response (RR) strategies. This strategy secures the information that depends on protection. This technique for the most part stays away from the bias answer. In 1965 Warner built up the strategy in which measure the individuals rate in the populace that has the particular attributes. In this strategy the respondent can't offer the mistaken response or they won't give the answer [9]. in the survey which include the sensitive information to reduce both response and non-response bias the randomized response technique is designed. To protect the privacy of response of a person it uses probability theory. Some sensitive research areas like drugs, assault and drugs use this type of techniques. To hide the real status of the respondent the data should scrambled in a way that the respondent could not distinguished [10]. To solve the problem of survey Warner used RR technique. in that approach Related and Unrelated-Question Model have been proposed.

a. Scheme used one group

The entire attribute create one group. By getting same value all attributes stay together. Here we take tree attributes A1, A2 and A3. There is probability of answering the questions of all the attributes is same. Either they tell truth or they tell lie.

$$P(A_1 = 1 \land A_2 = 1 \land A_3 = 0)$$

To present

$$P(A_1 = 0 \land A_2 = 0 \land A_3 = 1)$$

Involvement is come from P (001) and P (110), the derived equation is as follows [11].

$$P^*(110) = P(110).\theta + P(001).(1-\theta)$$

$$P^*(001) = P(001).\theta + P(110).(1-\theta)$$

To create the decision tree information required. The one group basic model is given below:

$$P^*(E) = P(E).\theta + P(\overline{E}).(1-\theta)$$

$$P^*(\overline{E}) = P(\overline{E}).\theta + P(E).(1-\theta)$$

The coefficiency matrix is:

$$\begin{pmatrix} p^*(E) \\ P^*(\overline{E}) \end{pmatrix} = \mathrm{M}_1 \begin{pmatrix} P(E) \\ P(\overline{E}) \end{pmatrix}, \text{ where } \mathrm{M}_1 = \begin{bmatrix} \theta & (1-\theta \\ 1-\theta & \theta \end{bmatrix}$$

b.Scheme used two group

Two group methods combined whole data in two groups. It enhances the security level because if the interviewer knows about the answer of one group then there is no probability of knowing the answer of other group. The two group model is as follows:

$$\begin{pmatrix} P^*(E_1\ E_2) \\ P^*(E_1 \overline{E_2} \\ P^*(\overline{E_1}\, E_2) \\ P^*(\overline{E_1}\, \overline{E_2}) \end{pmatrix} = M_2. \begin{pmatrix} P(E_1\ E_2) \\ P(E_1 \overline{E_2}) \\ P(\overline{E_1}\, E_2) \\ P(\overline{E_1}\, \overline{E_2}) \end{pmatrix}$$

The derived matrix is as follows:

$$\text{Where } M_2 = \begin{bmatrix} \theta^2 & \theta(1-\theta) & \theta(1-\theta) & (1-\theta)^2 \\ \theta(1-\theta) & \theta^2 & (1-\theta)^2 & \theta(1-\theta) \\ \theta(1-\theta) & (1-\theta)^2 & \theta^2 & \theta(1-\theta) \\ (1-\theta)^2 & \theta(1-\theta) & \theta(1-\theta) & \theta^2 \end{bmatrix}$$

c. Scheme used three group

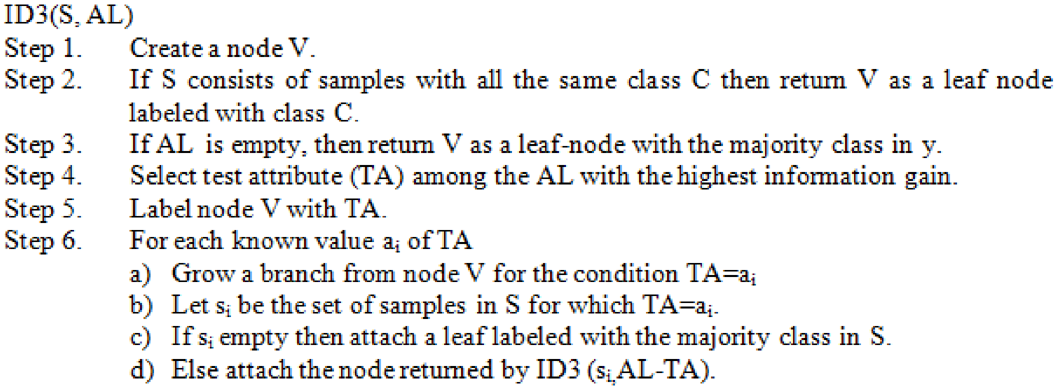
Three group methods divided the data in three groups in order to further enhance the privacy level. The model for the three-group is shown below:

$$\begin{pmatrix} P^*(E_1\, E_2\, E_3) \\ P^*(E_1 E_2\, \overline{E_3}) \\ P^*(E_1 \overline{E_2}\, E_3) \\ P^*(E_1 \overline{E_2}\, \overline{E_3}) \\ P^*(E_1 \overline{E_2}\, \overline{E_3}) \\ P^*(\overline{E_1} E_2 \overline{E_3}) \\ P^*(\overline{E_1}\, \overline{E_2}\, E_3) \\ P^*(\overline{E_1}\, \overline{E_2}\, \overline{E_3}) \end{pmatrix} = M_3 = \begin{pmatrix} P(E_1\, E_2\, E_3) \\ P(E_1 E_2\, \overline{E_3}) \\ P(E_1 \overline{E_2}\, E_3) \\ P(E_1 \overline{E_2}\, \overline{E_3}) \\ P(E_1 \overline{E_2}\, \overline{E_3}) \\ P(\overline{E_1} E_2 \overline{E_3}) \\ P(\overline{E_1}\, \overline{E_2}\, E_3) \\ P(\overline{E_1}\, \overline{E_2}\, \overline{E_3}) \end{pmatrix}$$

$$M_3 = \begin{bmatrix} \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta(1-\theta)^2 \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 & (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & (1-\theta)^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) \\ (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 \end{bmatrix}($$

Building Decision Trees

For classification of data decision tree is used. The algorithm is described below:



```
ID3(S, AL)
Step 1.   Create a node V.
Step 2.   If S consists of samples with all the same class C then return V as a leaf node
          labeled with class C.
Step 3.   If AL is empty, then return V as a leaf-node with the majority class in y.
Step 4.   Select test attribute (TA) among the AL with the highest information gain.
Step 5.   Label node V with TA.
Step 6.   For each known value $a_i$ of TA
          a) Grow a branch from node V for the condition TA=$a_i$
          b) Let $s_i$ be the set of samples in S for which TA=$a_i$.
          c) If $s_i$ empty then attach a leaf labeled with the majority class in S.
          d) Else attach the node returned by ID3 ($s_i$,AL-TA).
```

VIII. Result

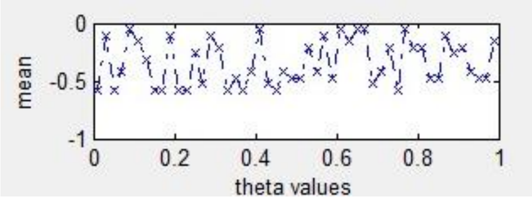


Figure 5: Mean Value of group 1

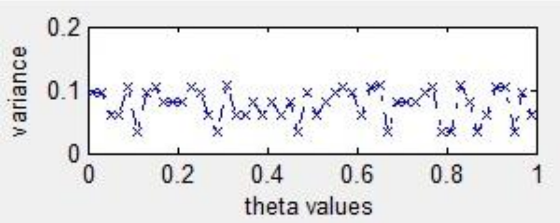


Figure 6: Varianee of group 1

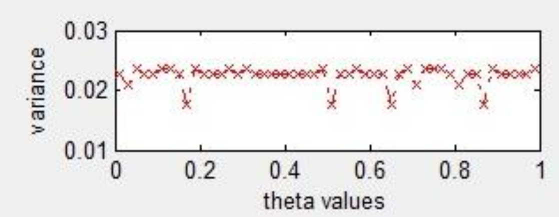


Figure 7: Variance of group 2

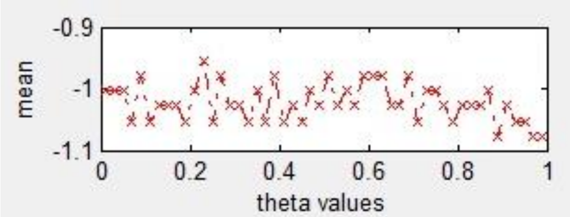


Figure 8: Mean Value of group 2

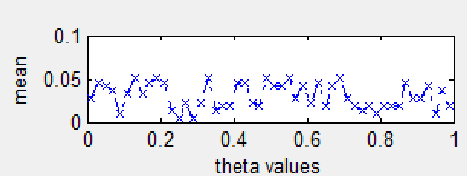


Figure 9: Mean Value of group 3

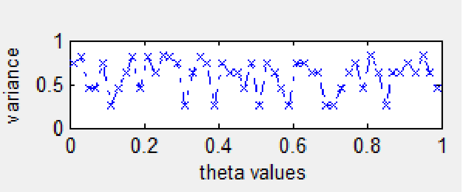


Figure 10: Variance of group 3

IX. Conclusion

In data mining process the private data preservation is very important. This is the research area where many researchers developed their research in several ways. To avoid the bias answer and to prevent the survey data it developed randomized response technique. Specific randomness add with the answers to protect the survey data. For increase the level of privacy the research work use one, two and three group methods. Data randomized in different way for each group. To classify the data the decision tree used which depends on these groups.

REFERENCES

1. Monika Soni, Vishal Shrivastva, "Randomized Response Technique in Data Mining", International Journal on Recent and Innovation Trends in Computing and Communication,Volume: 1 Issue: 6.
2. Pallavi Wankhade, Prof.R.R. Shelke, "A Study of Data Mining Tools in Knowledge Discovery Process",International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue III, March 2016.
3. “Data Mining: An Introduction”, SPSS Whitepaper. SPSS. 2000.
4. Deependra Dwivedi, “Study Analysis of data mining Algorithms: case study” Researcher. 2012;4(2):16-19] 2012, http://www.sciencepub.net.
5. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, “A study of data mining tools in knowledge discovery process”, IJSCE, Volume-2, Issue-3, July 2012.
6. An Overview of Data Mining Techniques Excerpted from the book by Alex Berson, Stephen Smith, and Kurt Thearling. Page no 2.
7. Lior Rokach and OdedMaimon, “Top-Down Induction of Decision Trees Classifiers – A Survey” IEEE Transactions On Systems, Man And Cybernetics: Part C, Vol. 1, No. 11, November 2002.
8. Carlos N. Bouza1, Carmelo Herrera, Pasha G. Mitra,”A Review Of Randomized Responses Procedures The Qualitative Variable Case”, Revista Investigacion Operacional VOL., 31 , No. 3, 240-247 2010.
9. Zhouxuan Teng, Wenliang Du,”A Hybrid Multi-Group Privacy-Preserving Approach for Building Decision Trees”.
   Gerty J. L. M. Lensvelt-Mulders, Joop J. Hox And Peter G. M. Van Der Heijden “How To Improve The Efficiency of Randomised Response Designs “, Springer 2005.