Indian Journal of Science and Technology, *Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/86386, January 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Enhancing Privacy Preservation in Data Mining using Cluster based Greedy Method in Hierarchical Approach

R. Hariharan*, C. Mahesh, P. Prasenna and R. Vinoth Kumar

Department of Information Technology, Vel Tech University, Chennai - 600062, Tamil Nadu, India;
hharanbtech@gmail.com, chimahesh@gmail.com, prassennacse@gmail.com, vinothtechnocrat@gmail.com

Abstract

Background/Objectives: Privacy preservation in data mining to hold back the sensitive data from attackers. **Findings:** There are various existing methods available to preserve the data like perturbation, anonymization, randomization etc., each method has its own advantages and disadvantages. The trade-off between security and utility of data should be handled with standardizing methods for the PPDM. In this paper explained a method based on PPDM in data mining using cluster based greedy method. **Application/Improvements:** This method can be applied in sensitive data areas such as hospitals, Customer Management System, government survey, etc., where there is need for privacy preservation.

Keywords: Cluster based Greedy Method, Classification Error, Isometric Transformation, Privacy Preservation, Privacy Preservation Rate

1. Introduction

Data mining affects the privacy of the individual. However, the data when altered to preserve this privacy affects the results of data mining. This led emergence of interesting research area called privacy preservation data mining. There are various techniques proposed earlier like randomization, k-anonymity.

The research has been started in 1991 in the area of PPDM. The work can be categorized into two major types[6] - perturbation and anonymization. Perturbation is a mathematical method that gives approximate solution to the problem which cannot be solved exactly. Data perturbation is described as protection of data in sensitive case using some mathematical application. Data anonymization is the process of destroying tracks on the data that would lead an intruder to its origins. With few evaluations, we can conclude that anonymization is the effective for preserving privacy.

Hospital data base management is one of the main examples for PPDM. When patient's data are given for mining, even after removing unique attributes like name, patient ID etc., but still the particular record can be identified with address, zip code, age etc. Credit card details management also requires PPDM. The person should buy their needed things by using their credit card. If we hack their credit card information we can easily find what are the things is purchased especially when some products is treated as sensitive. Hence there is a need to hide those information from that of unauthorized persons by using PPDM.

In the government sector also people's details are stored. Other than the name, all other information are very sensitive for example in passport the name can be

**Author for correspondence*

shown but the age, address, the place where he/she visited all are consider as sensitive; so we have to keep that data as private.

2. Categorization of PPDM

PPDM techniques can be categorized based on the following five areas:

- Data distribution- referring the data are centralized or distributed.
- Data modification- referring modification of data values to ensure the privacy like aggregation and swapping.
- Data mining algorithm referring the target of DM algorithm for PPDM method is defined.
- Data or rule hiding for check the PPDM method hide the raw or aggregate data.
- Privacy preserving technique that is used PPDM. Heuristic, cryptography, reconstructing based.

3. Related Work

In[1], classified privacy preserving data mining techniques, including data modification and cryptographic, statistical, query auditing and perturbation-based strategies. In[2] first represented the attributes in numeric format according to the distribution based on domain generalization hierarchy and extended by setting the restriction of the valid generalization. Multiplicative perturbation was used in[3] where the original data of each data provider are multiplied with the same matrix which is random and orthogonal before released, however this kind of perturbation was easily reconstructed by methods such as Principal Component Analysis (PCA), i.e., recovering the original data by analyzing the covariance matrix of the perturbed data.

In[4] a set of hybrid transformations has been introduced to ensure privacy of categorical data in clustering. The misclassification errors obtained after applying the hybrid data transformation techniques for various noise levels are computed and they are found to be the least for a noise level of 75. The method is specialized for categorical data. The misclassification errors are comparatively more if the categorical attributes are altered. The paper[5] specifies the methodology % of perturbation by random projection technique. Using this method the quality of data is disturbed and the procedure is irreversible. Successful reconstructions essentially mean the leakage of privacy, so this work identifies the possible risks of RP when it is used for data perturbations. In[6] converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. An accurate decision tree can be built directly from those unreal data sets. This novel approach can be applied directly to the data storage as soon as the first sample is collected.

In[7] proposes a technique which is computationally more expensive than some comparable methods. The author had performed cluster displacement followed by cluster rotation. The disadvantage of this method is, due to huge amount of data processing it leads to misclassification error. The computation time is also high compared to general clustering methods. However, these additional computational costs are effectively offset by the increased security measure offered by this method through independent rotation of clusters to perturb the data. In[8] proposes a framework to transform each longitudinal patient record into a form that is indistinguishable from at least k−1 other records. This is achieved by iteratively clustering records and applying generalization, which replaces ICD codes and age values with more general values, and suppression, which removes ICD codes and age values. This method performs better but does not take into account of homogeneous attack. Also there is a chance that the statistical values of the data are altered completely. The method also lags in data reclassification, since the data are generalized without any supervision.

Therefore, there is requirement which assimilate the PPDM procedures, that overcomes the various disadvantages like information loss, homogeneous reidentification, statistical similarities etc., without compromising the data privacy. In this paper, we are discussing about an anonymization using Nested Clustering. The method performs better on all the discussed issues of existing methods and provides efficient privacy of data.

4. Existing Methodologies

4.1 Additive Perturbation

This method[9] is applicable for attributes one by one only. In this case the randomized noise (A) is added to the set of records (B). The summation of entire random values should be zero ($\Sigma A=0$). The random values are added to

the original data set value and saved in the new record set (C). That new output dataset values are positive and are nearest to the original data, hence the new data set seems like original data set. The methodology is simple but does not take care of mining results or correlation among the attributes. The method is also reversible to a certain extent.

4.2 Multiplication Perturbation

In multiplication perturbation[9] is defined as adding of noise (X) with original value (Y), noise is nothing but random values with the product value is equal to one. Then that random values are multiply with original record set. That stored in new data set (Z), values should not be negative. Z=X*Y the X value is very small value. After multiply the values, need to check with parameter value with original value.

4.3 Isometric Transformation

Isometric transformation Equation (1) is one of the special geometric transformation. The main character of the geometric transformation is preserving objects are moving among them in n- dimension Euclidean space. In Euclidean space the distance must be invariant. T is isometric transformation if it preserve the distance it have to satisfy the condition $|T(p) - T(q)| = |p - q|$ for all p, q € R^n.

The various operations of isometric transformation is

- Translation is defined as shift the position of data in constant distance in that parallel direction.
- Rotation - Which have a center value $|T(p)-a| = |p-a|$ for all p.
- Reflection it mark all the points in mirror image in a (d-1) fixed dimension.

In this process mainly focus on rotation on 2D discrete space, here perform rotation based on coordinate axis. The rotation in 2D discrete space. After that to perform the rotation in 2D discrete space by measuring the angle Ø using the below formula (R), the rotation is clock wised in x and y co-ordinates. So the rotation of 2D discrete space could refer as matrix form v"=RV here R is 2×2 rotation matrix. V is vector value having the real dataset and V" is the column value are rotated coordinates .

$$R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad \text{Equation (1)}$$

4.4 K-Anonymity Transformation

The main purpose of this k-anonymity[10] algorithm is capable of transforming a nonanonymous dataset into a k-anonymity[12] data set. Here the entire operation done by using classification tree adult set. In The adult data set[11] assuming k-anonymity = 100. Anonimization operation[12] can be performed successfully with the k- anonymity classification tree algorithm. In this classification tree, apply the algorithm to the nodes with height =1[9]. Node 9 and 19 have height = 1, so choose the node 9. The children (nodes 10-14) are examined indicates that only node 11 complies with k (because 190 > 100). The remaining siblings (nodes 10, 12, 13, and 14) have 6 +3 +1 + 2 = 12 instances in total, i.e., there are 88 instances missing. Node 11 has 90 extra instances (190 -100 = 90). Thus, 88 instances of node 11 can be randomly selected and used to complement the uncompliant nodes. The remaining 102 instances of node 11 can be mounted as the anonymous data set with the following quasi-identifiers revealed (all non-quasi-identifier attributes, such as hours per-week, can be revealed as well): marital-status = Married-civ-spouse, education = Some- college, occupation = Exec-managerial, work class = Private, race = White. Finally, node 9 is prune resulted with the following revised tree presented in the given new tree, the next node which is to be examined is node 19.

5. Metrics

Some of the metrics[13,14] need to check calculate the preservation of data.

- Loss of data is use to check how much amount of data is lost can be found with the following steps
  Step 1: Find the centroid value.
  Step2: And subtract the centroid value from the original data.
  Step3: Sum the total row value.
- Discloser measure is calculate using two metrics

a. Bias in mean Equation (2) and
b. Bias in standard deviation Equation (3)

$$BIM = (\bar{Y} - \bar{X}) / \bar{X} \qquad \text{Equation (2)}$$

$$BISD = (sY - sX)/sX \qquad \text{Equation (3)}$$

X is mean of original data and y is mean of altered data. Sx and sy is standard deviation of original data and perturbed data respectively.

- Classification of error rate shows the amount of error is present and it should be minimum
- Computational time is nothing but execution time of the process, how many times the process executes and the time required.
- Privacy preservation rate is explained as the number of altered data divided by the number of original data. It is easy find the count of unpreserved data and find the percentage.
- Measure of privacy is defined as variance of difference between the original and modified data that is divided by variance of original data.
- Regression is defined as the sum of difference between original data and modified data divided by total number of data.
- Rand Index Equation (3) well-known measure of agreement called the Rand index was used to measure the similarity between two data clusters. To compare two partitions X and Y of a given set of elements S, the Rand index is given by the following: where a represents the number of true positives, i.e., the number of pairs S in that belong to the same set in both X and Y. The variable b represents the number of true negatives, i.e., the number of pairs in S that belong to different sets in both X and Y. The numerator denotes the number of agreements between the two clusters.

The variable c represents the number of false positives, i.e., the number of pairs in S that belong to the same set in X and different sets in Y. The variable d represents the number of false negatives, i.e., the number of pairs in S that belong to the same set in Y and different sets in X.

The summation of c and d represents the number of disagreements between two clusters, and thus, the denominator represents the total number of decisions. The range for the Rand index value is 0 to 1. A higher value for the Rand index implies higher similarity, and a value of 0 indicates that there is complete disagreement between the two clusters.

$$RI = \frac{a+b}{a+b+c+d} \qquad \text{Equation(4)}$$

6. PPDM (PPDM in Data Mining using Cluster based Greedy Method)

In this paper about to see how the drawbacks of existing system is rectified using new method, here the proposed method is called "PPDM in data mining using cluster based greedy method". First choose the database with sensitive data and remove the unique identify attribute, after that it will be cluster that database using any one of the main attribute using FCM algorithm in mat lab. After that main cluster is divided in to sub cluster and rotate that sub cluster's data for preserving purpose. One of the main problem is how is numbering to the clusters, because the sub cluster are spread here and their randomly. It is not easy to numbering, so apply greedy method. The greedy method is nothing but first calculate the centroid of the main cluster and need to calculate the mean for all sub cluster using the mean value and calculate the sub cluster1 (s1) which centroid is very nearer to the main cluster centroid and marked as visited, after that calculate the distance between centroid of s1 to all the sub cluster's centroid after that find the shortest distance and marked as sub cluster (s2) which have very shortest distance with s1 and mark it as visited after that continue this operation till the last cluster is visited. Once the cluster is numbered that is consider as visited, so in next calculation no need to consider that visited cluster. After that perform the rotation in that sub cluster.

The data from the sub cluster 1 are move towards centroid of the sub cluster s2 like the data's are modified till the last cluster. In this operation the datas are preserved to calculate the metrics value for this operation. For PPDM this paper Implement "PPDM in data mining using cluster based greedy method" is most effective solution for preserving the data's comparing with the existing technique. Here preserving the data using rotation method in nested clustering. First choose the sub-cluster (s1) cluster which is very near to centroid and from that cluster calculate the distance of nearer cluster find the next cluster which have shortest distance and number as s2 and from s2 need to calculate the shortest distance to nearer cluster which is not visited. Continue this operation till the last cluster is visited. After that the rotation operation is performed, in that rotation the data of s1 is moved towards s2 centroid, like perform this operation. After this operation need to calculate metrics is define in this paper. After operation, the data is modified. So sensitive data is not leaked to unauthorized users. Huge amount of data sufficiently used in this method. Here time complexity and also very less comparing to existing system.

6.1 Experimental Setup

In our case is implemented for the Adult dataset, which contains 32561 records, of which 30722 are complete.

There are 14 attributes in the data base[6], of which we have taken {Age, Work-class, Education, Hours/week, sex, race, Marital-status, Salary}. {Salary} is considered as the sensitive attribute and others as Quasi Identifiers. In QI {Age} and {Hours/week} attributes are numeric and others are categorical. The numerical attributes are only anonymized leaving the categorical values as they are used. This experiment is done in uci repository in MATLAB.

6.2 Flow Diagram

The flow diagram Figure 1 is shows the process of our greedy method which retrieve the data set from the data base here the adult data base as an example. The data pre-processing is nothing but arrange the data in some standard order then need to remove the unique attribute which is easy to identify the person. for example phone number and address. Using this attribute unauthorised can easily identify the person so must remove that attribute. After that cluster the data. In this paper using the FCM algorithm to cluster the data. In that cluster if it's rotate the unauthorized person can easily find the data, so cluster that main cluster in to sub-clusters then need to number the all sub-cluster using the greedy method. Before that find the sub-cluster1so calculate the centroid of main cluster and calculate centroid for sub-cluster also, from that centroid of the sub-cluster value need to calculate the mean from that values need to find the shortest distance ,the sub-cluster which is have very small value is marked as sub-cluster1 (s1) after that the paper is proposed greedy method. The greedy method is nothing but calculate the shortest distance from the centroid of s1 to all others sub-cluster centroid. Then compare all the distance value and select the sub-cluster which is have small value and marked as visited, once visit the sub-cluster then it never consider that sub-cluster for future calculation. This operation is continue till last node is visited. After that perform the rotation in that rotation the elements of sub cluster is moved towards the centroid of s2 like that all elements are moved to its adjacent cluster centroid. Now perform the validation now that data base with privacy preserved for data mining.

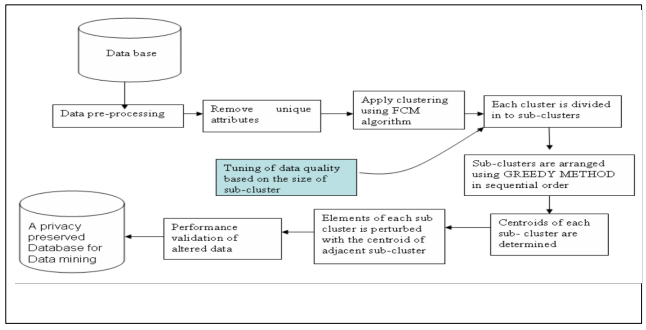

Figure 1. Greedy method.

6.3 Algorithm

From the algorithm is implemented for the greedy method for PPDM. That algorithm shows the step by step process of our method. First use the FCM method for getting the cluster and sub cluster. Calculate centroid of main cluster and sub-cluster. From that identify the sub-cluster which is near to centroid of main cluster. Mark that sub- cluster as s-1 and calculate near sub-cluster's Euclidean distance using greedy method. Mark sub-cluster2 which is have shortest distance from centroid of s1, like that continue this operation. After that, perform the rotation. The elements of s1 move towards s2. Like that this operation is repeated till the last node is visited. This Figure 2 is a simple example of our method. Consider the whole circle is main cluster and the small circles are sub-clusters the black dot is centroid of the main cluster and the red dot is centroid of the sub-cluster, the sub-cluster1 which was chosen is very nearest to centroid.

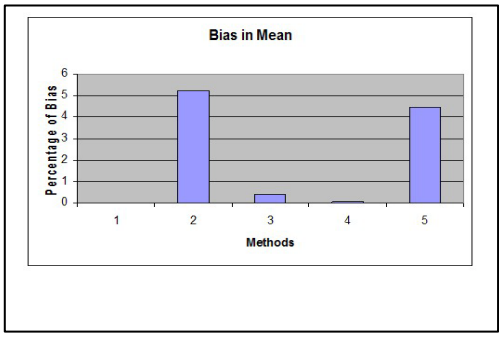

Figure 2. Proposed work

Steps to implement rotation after greedy method for sequencing the sub-clusters

- Apply fcm procedure to the whole data and find idx with vaues1/2/3.
- Separate the three clusters .
- Take only the elements of cluster1.
- Apply fcm to cluster1 data and by max() function find idx values 1/2/3/…30.
- [C,I] = max(...) finds the indices of the maximum values of A, and returns them in output vector I. If there

are several identical maximum values, the index of the first one found is returned.

- Apply the greedy sequencing algorithm.
- for I = 1 to n//n = number of records ic cluster1 idx(i) = rank(idx(i)).

//greedy method for sequencing the sub-clusters Input: Set of all sub-clusters with their centroids Output: Sequential order of the given sub-clusters Method: //C = Centroid of the chosen cluster //ci = centroid of the chosen sub-cluster //m = Number of sub-clusters // Euclidean distance b/w the centroid C and a sub-cluster centroid ci d(C, ci) = sqrt[(Cage - ci, age)2 + (Chpw - ci,hpw)2]

```
1.      for ( i=1 to m)
        rank[i]=0
2.      min=0 , rank=1
3.      for(i=1 to m)
a.      if (rank[i]=0)
(i)     if (d(C, ci ) < min )
   →    min= d(C, ci )
   →    minpos=i
4.      rank[minpos]=j
```

6.4 Comparison with Existing Methods

From the Table 1 shows the comparison of metrics values in greedy method and with other existing methods.

In the Table first column is represent the metrics. Second column is represent units of the metrics. Third column contains original data. Column 4, 5, 6, 7, 8, 9 having a metrics value for additive perturbation, multiplication perturbation, isometric transformation, K-anonymity transformation, cluster rotation and greedy based sequence cluster rotation respectively. The following graphs are shown the performance of original data and existing method and the greedy method Performance:

In the graphs the numbers 1, 2, 3, 4, 5 are specific methodology.

- Additive perturbation.
- Multiplicative perturbation.
- Isometric transformation.
- K-anonymity.
- Greedy method.

Comparing the above technologies from the the graph (Figure 3. Loss of data, Figure 4. Bias in mean, Figure 5. Classsification error, Figure 6. Privacy preservtion rate). The isometric transformation has given the best result in all metrics. All elements from the sub-cluster is moved to the nearest sub-cluster with particular angle only, so the data's are modified evenly and it moved inside the main cluster only. But that method is fully reversible, if the unauthorized person know the angle value the then they will do the reversible operation and he can easily find the original data. So we have to compare other methodologies. The next best method is our greedy method. In our method the elements from one sub-cluster are moved to another sub-cluster which have shortest distance to each other. So the data's are modified evenly and the data's are moved inside the main cluster only. And calculating all metrics is also supporting to the greedy method. For example the information loss is very small and the rate of classification error also very small, and its one time computational only, the maximum data is modified so it give the better preservation rate. So comparing to above technologies the greedy method is given the best preservation and secure and easy to implementation.

Table 1. SSGM comparison with existing methodology;

Parameters	Units	original data	AP	MP	IR	KA	CR	GBCR
Loss of Data	Cont.	20.399	50.525	16.481	17.857	57.342	20.591	16.334
Bias in Mean	Cont.	39.68	39.68	41.76	39.51	39.7	37.91	28.45
Bias in Standard deviation	Cont.	15.408	18.947	33.555	15.332	14.562	11.589	9.542
Rate of Classification Error	%	189.47	13.2	15.7	0	9.2	14	8.9
Computational time	Sec	Nil	O(2n)	O(2n)	O(n)	O(n)+1	O(n)+2	O(n)
Regression	Cont.	Nil	0	0.009	0.001	0.002	0.002	0.001
Privacy preservation rate	%	Nil	0.041	0.038	0	0.163	0.254	0.342
Rand Index	%	Nil	11.3	16.4	9.1	7.6	12.4	12.2
Measure of Privacy	Cont.	Nil	35.37	15.194	2.43	8.82	35.96	36.25

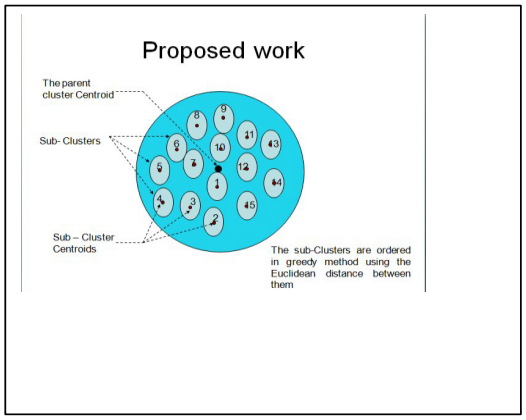

Figure 3. Loss of data quality

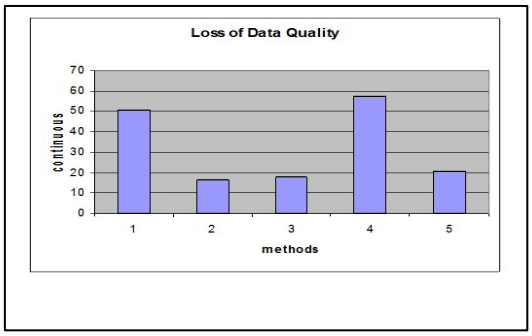

Figure 4. Bias in Mean

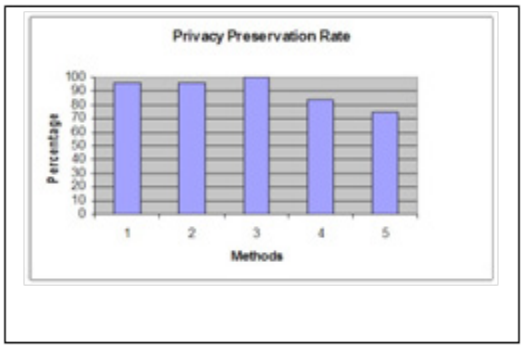

Figure 5. Classsification

Figure 6. Privacy preservtion rate

7. Conclusion

In this paper, the proposed new method is greedy method for privacy preservation Data mining. This paper metrics compares the several of the existing methods and greedy method. We conclude that the greedy method is the best technology for PPDM with secure manner. In future, the method can be enhanced by using a parameter for rotation which depends on the distance between the sub-cluster to improve the quality of data as well as the preservation rate.

8. References

1. Aggarwal C, Yu P. Privacy-preserving data mining. Models and Algorithms. Springer: Berlin Heidelberg; 2008.
2. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002 Oct; 10(05):571–88.
3. Oliveira SRM, Zaiane OR. A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Computers and Security. 2007; 26(1) :81–93.
4. Rajalaxmi RR, Natarajan AM. An effective data transformation approach for privacy preserving clustering. Journal of Computer Science. 2008; 4(4):320–26.
5. Sang Y, Shen H, Tian H. Effective reconstruction of data perturbed by random projections. IEEE Transactions on Computers. 2012 Jan; 61(1):101–17.
6. Fong PK, Jahnke JH. Privacy preserving decision tree learning using unrealized data sets. IEEE Transactions on Knowledge and Data Engineering. 2012 Feb; 24(2):353–64.
7. Shivaji S, Ameer DM, Khan A, Khan W, Challagalla A. Privacy preservation in k-means clustering by cluster rotation. IEEE Region 10 Conference TENCON'09; Singapore. 2009. p. 1–7.
8. Tamersoy A, Loukides G, Nergiz MEN, Saygin Y, Malin B. Anonymization of longitudinal electronic medical records. IEEE Transaction on Information Technology in Biomedicine. 2012 May; 16(3):413–23.
9. Li XB, Sarkar S. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Transactions on Knowledge and Data Engineering. 2006 Sep; 18(9):1278–83.
10. Chiu CC, Tsai CY. A*k*-anonymity clustering method for effective data privacy preservation, Advanced Data Mining and Applications, Springer-Verlag: Berlin Heidelberg; 2007. p. 89–99.
11. Silevich S, Rokach L, Elovici Y. Efficient multidimensional suppression for K-anonymity. IEEE Transactions on Knowledge and Data Engineering. 2010 Mar; 22(3):334–47.

12. Wu Y, Sun Z, Wang X. Privacy preserving k-anonymity for re-publication of incremental datasets. World Congress on Computer Science and Information Engineering; 2009. p. 53–60.
13. Rajalakshmi V, Mala GSA. Anonymization based on nested clustering for privacy preservation in data mining. Indian Journal of Computer Science and Engineering, 2013 Jun-Jul; 4(3):216–24.
14. Nagarajan S, Chandrasekaran RM. Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques. Indian Journal of Science and Technology. 2015 Apr; 8(8):771–6.