ANONYMIZATION BASED ON NESTED CLUSTERING FOR PRIVACY PRESERVATION IN DATA MINING

V.Rajalakshmi
Research Scholar
Sathyabama University
Chennai, India
rajalakshmi.bala03@gmail.com

G.S.Anandha Mala
Professor & Head,
Dept. of CSE, St.Joseph's College of Engg.
Chennai, India
gs.anandhamala@gmail.com

Abstract

Privacy Preservation in data mining protects the data from revealing unauthorized extraction of information. Data Anonymization techniques implement this by modifying the data, so that the original values cannot be acquired easily. Perturbation techniques are variedly used which will greatly affect the quality of data, since there is a trade-off between privacy preservation and information loss which will subsequently affect the result of data mining. The method that is proposed in this paper is based on nested clustering of data and perturbation on each cluster. The size of clusters is kept optimal to reduce the information loss. The paper explains the methodology, implementation and results of nested clustering. Various metrics are also provided to explicate that this method overcomes the disadvantages of other perturbation methods.

***Keywords*:** Nested Clustering, Anonymization, Privacy preservation, Perturbation.

1. Introduction

The technology development in storage and processing increased the amount of data storage and processing. Data mining has become a necessary task today in all fields to extract the hidden useful information. This knowledge is used for improving our quality of life and getting known about the facts which are not seen obviously. On the one hand, such data is an important asset to business organizations and governments for decision-making processes and to provide social benefits, such as medical research, crime reduction, national security, etc. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly. Data mining has become a necessary task today in all fields to extract the hidden useful information[4]. This knowledge is used for improving our quality of life and getting known about the facts which are not seen obviously. But the mining operation in addition to extract the useful information, since it has to operate on the real world data the privacy of the individual, owning the data is highly affected.[2,14].

A dataset is having three kinds of attributes. The individual specific attributes like Name, SSN Number, Bank account number, etc., are removed as they are not of use in mining. The next set of attributes called quasi identifiers which can be collectively used for identifying the records. The remaining attributes are called sensitive attributes which have to be protected for security [3,14]. The task of privacy preserving algorithms is to protect the identification of sensitive attribute value for the corresponding quasi identifier attributes. Many of the Privacy preserving algorithms alter the quasi identifiers in order to make them not to identify the particular sensitive data.

PPDM techniques can be broadly classified into two types. One is Perturbation /Anonymization methods – which alter the data by suppression, generalization, additive or multiplicative factor, fuzzy based [8], random number or geometric projections [5]. Perturbation methods are mainly used with a compromise on data utility, as the data are altered and or not reversible. But the privacy is provided to an extent except closeness attack. The most popular algorithms for the data mining research community in this category are k-anonymity and ℓ-diversity. K-anonymity requires each tuple in the published table to be indistinguishable from at least k-1 other tuples. Tuples with the same or close QI values form an equivalence class [15]. However, k-anonymity cannot protect against homogeneity and background knowledge attacks. To address these

shortcomings, the ℓ-diversity principle was proposed, which requires that different values of the sensitive attributes are well represented in each equivalence class, thus preventing an attacker from guessing the sensitive attribute value for a QI set with probability greater than 1/ℓ for each equivalence class.

The other one is cryptographic based method – they use a public or private key to hide the data and reconstructed when required. The privacy is role based and executed with the level of security of the key values. The privacy preservation is expected to provide the maximum protection of data with minimum loss.

In this paper we have implemented Anonymization based on nested clustering algorithm. We have implemented for the Adult dataset, which contains 32561 records, of which 30722 are complete. There are 14 attributes for the data base, 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.

The architecture shows the two phases of our procedure as

1. Nested clustering phase
2. Anonymization phase

In the Nested clustering phase, the original database is clustered efficiently into enough number of sub clusters, each having a minimum of 3 records and a maximum of 5 records in them. This is done by grouping and re-clustering repeatedly. The sub clusters are numbered sequentially for their usage in the next phase.

In the Anonymization phase, the numeric values are moved towards the centroid of each of the sub clusters. If the centroid of a sub cluster matches with its value, then the record is moved towards the centroid of the parent cluster. Thus the objects are made to remain in the same cluster with their values perturbed.

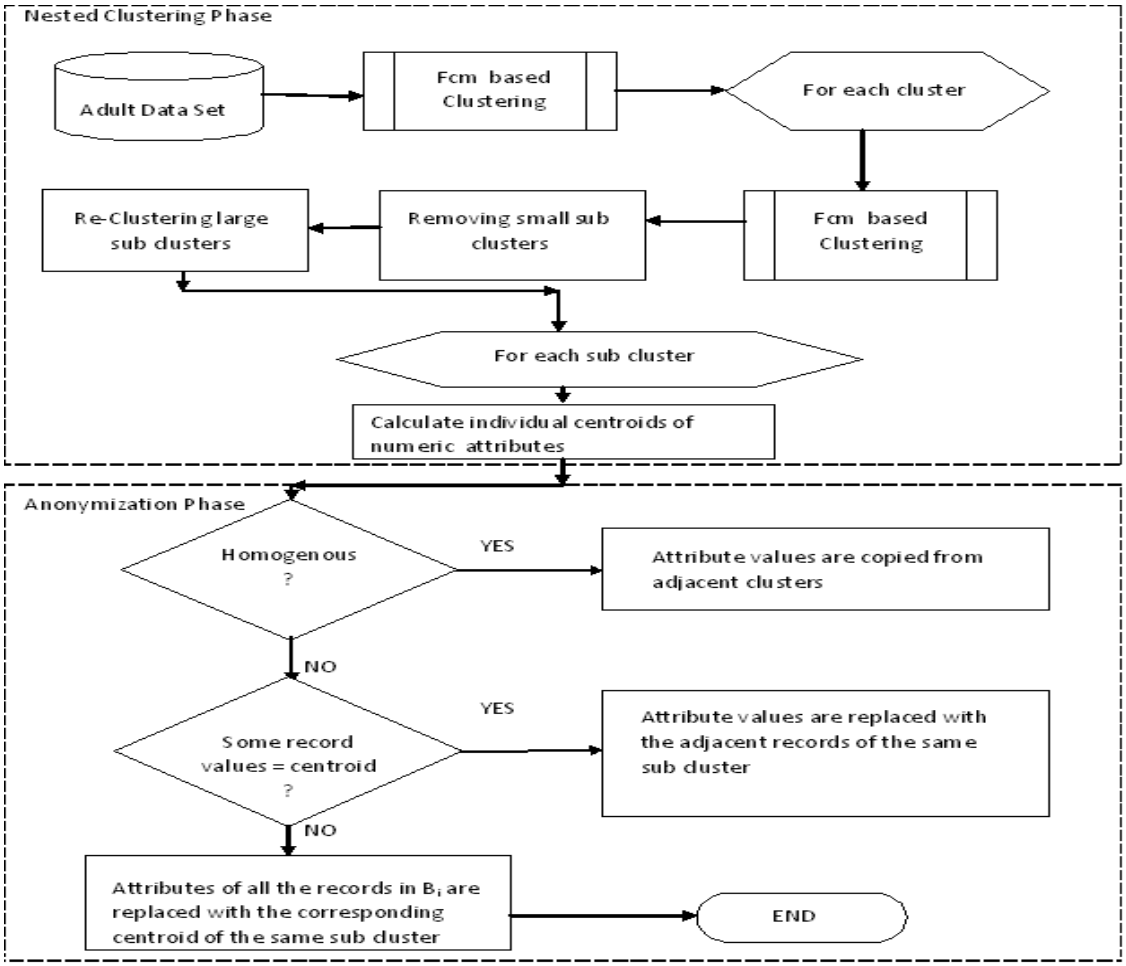

Figure 1: Flow Diagram of ANC

2. Related Work

In [2], Aggarwal and Yu classify privacy preserving data mining techniques, including data modification and cryptographic, statistical, query auditing and perturbation-based strategies. In [13] Sweeny 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 [9] 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 analysing the covariance matrix of the perturbed data.

In [11] 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 [16] 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.

[10] 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.

[12] 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.

[1] 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 re-classification, 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 (ANC). The method performs better on all the discussed issues of existing methods and provides efficient privacy of data.

3. Problem Definition

The above literatures revealed that there are multiple works done towards privacy preservation. But an algorithm with less mis-classification error and Anonymization done to each and every object is lacking. This lead to the development of our work Anonymization based on Nested Clustering (ANC).

4. Anonymization using Nested Clustering (ANC)

4.1 Definitions

Definition 1 : Sensitive item s -. The set $S \varepsilon D$ of items that represent a privacy threat if associated to a certain transaction constitutes the sensitive items set

$$S = \{s_1 \ldots ; s_m\}, m = |S|. \quad (1)$$

Definition 2: A quasi-identifier of table T, denoted as QI, is a set of attributes in D, which is externally available and can be exploited for linking to re-identify individual records with a significant probability. Potentially, any non sensitive item is a quasi-identifier.

$$QI = \{q_1, \ldots, q_{d-m}\} \quad (2)$$

$$D = \text{item set and } d = |D| \quad (3)$$

Definition 3: Privacy-Preserving Transformation- A transformation (or a function) T is privacy preserving if, for any data set D, the composition transformation T −1(T (D)) does not give D back, i.e.,

$$T^{-1}(T(D)) != D. \quad (4)$$

Definition 4: Divisive Hierarchical Clustering - Clustering is the task of grouping a set of objects in such a way that objects in the same group called cluster, are more similar to each other than to those in other groups. Hierarchical clustering is a set of nested clusters that can be organized as a tree. Each node is the union of its children and the root of the tree is the cluster containing all the objects.

If we start with one, all inclusive cluster and at each step, split a cluster until singleton clusters of individual points remain, is called Divisive Hierarchical Clustering.

Definition 5: Davies–Bouldin index - It is used to assess the quality of clustering algorithms based on internal criterion The Davies–Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma i + \sigma j}{d(ci,cj)}\right) \quad (5)$$

Where n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of all elements in cluster x to centroid c_x, and d(ci,cj) is the distance between centroids c_i and c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm.

4.2 Procedure summary

The database is clustered into two clusters with the salary values "<50" and ">=50". These two clusters are then handled separately. Each of them is sub clustered by

$$\text{No. of sub clusters} = \text{No. of records within the cluster}/3 \quad (6)$$

Each of these sub clusters are analyzed if they are having enough records (3) to anonymize. If it is less, they are merged with the adjacent clusters. Now all the clusters will have a minimum of three records.

The maximum number of records in a sub cluster needs to be decided, because in a greater sized cluster, if the attribute values are replaced by the centroid values, the amount of information loss will be more. Therefore, the size of the sub clusters should be restricted to 5. If the size exceeds, those sub clusters are clustered again till the number of each sub cluster remains less than 5. These biased sub clusters are anonymized efficiently to fit the criteria of privacy, efficiency and information loss.

4.3 Algorithm

4.3.1Nested Clustering

Input: Original database D[a1,a2,…an].

Output: Biased sub clusters B[a1,a2,…an] arranged sequentially.

Method:

1. The adult database with 30722 records each with 7 attribute values are clustered into 2 clusters using Fuzzy C-means Clustering

(i)The membership values are used to categorize the records into two clusters – C1 and C2

(ii)For each cluster

(a) Size = number of records in the cluster
(b) SC = Size/3
(c) The records are clustered into SC number of sub clusters and are numbered sequentially according to its position.
(d) Setting the minimum size of a sub cluster - The sub clusters which have less than 3 records are grouped with the adjacent sub clusters.
(e) Setting the maximum size of a sub cluster – The sub clusters which have records more than 5 are reclustered into 2 or more sub clusters and the sequence numbers are reordered.

4.3.2Anonymization

Input: Biased sub clusters B

Output: Privacy preserved database B

Method:

1. For each biased sub cluster Bi
   (i) Calculate the centroid of all the numeric attributes of Bi as C_{age}, C_{hpw}
   (ii) If all records in B_i are homogenous
      (a) {age} and {hpw} attributes of all the records in Bi are replaced with the values towards the corresponding centroid of parent cluster.

   Else if some records in position j of Bi are equal to C_{age}, C_{hpw}
      (a) A midpoint between the centroid of parent cluster and Bi is identified.
      (b) {age} and {hpw} attributes of those records in Bi are replaced with the values towards that mid point.

   Else
      (a) {age} and {hpw} attributes of all the records in Bi are replaced with the with the values towards the corresponding centroid of the same sub cluster.
   (iii) After the replacement the statistical parameters are checked if they remain the same, else the values are updated.

5. Issues solved in ANC

(1) The sizes of the sub clusters are random between 0 to n. If the size is small, it is more vulnerable to attacks. On the other hand if the cluster size is large, the generalization may lead to information loss. Hence after sub clustering, the child clusters are regrouped by biasing. The smaller sub clusters are grouped with the adjacent ones. The bigger sub clusters are re-clustered into smaller ones. The size of the sub clusters are kept optimal from 3 to 5, which will resist from the above disadvantages.

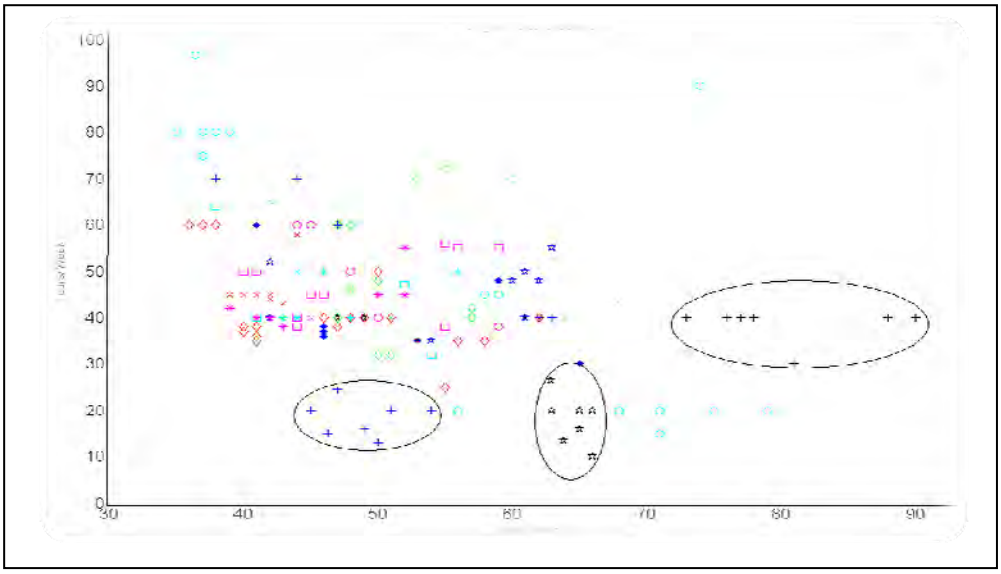

Figure 2: Original Sub clusters

The above plot shows the scattering of objects and sub-clusters are formed randomly. Some of the sub-clusters have many members (< 5) and some have very few (>3). If the numbers of members are more, the anonymity will result in more errors. Hence they are re-clustered into smaller clusters. If the numbers of cluster members are very less, then it is not possible to anonymise it by finding its centroid. Hence they are merged with its adjacent sub-clusters.

Figure 3 shows the biased sub- clusters, in which each of the sub clusters have a minimum of 3 objects and a maximum of 5 objects. Using such sub- clusters, the centroid is calculated and compared with each of its members.

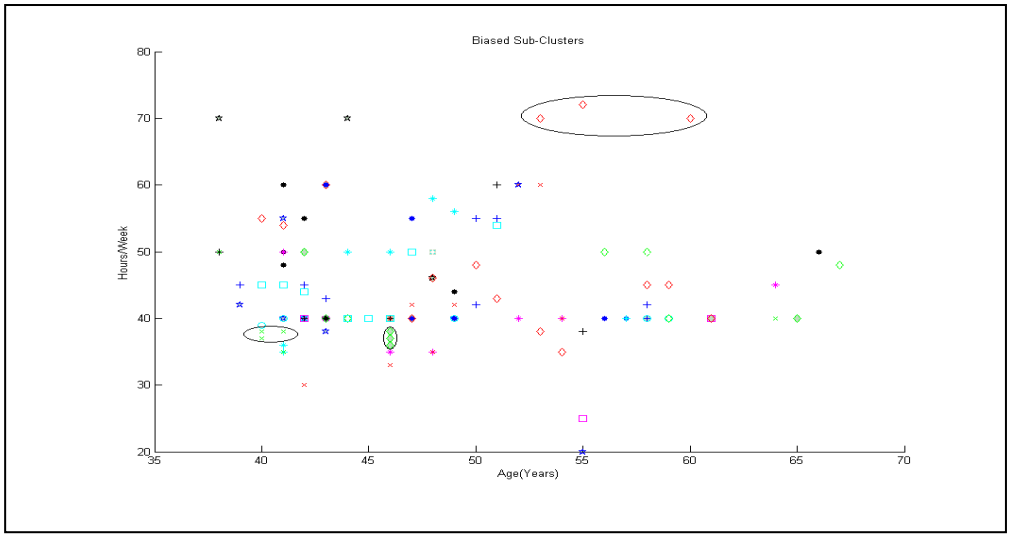

Figure 3: Biased Sub clusters

(2) If a whole sub-cluster is homogenous the replacement values have to be calculated separately. Since we know which is the adjacent cluster the centroid values of adjacent clusters are replaced. Thus we handle the K-Anonymity problem.

(3) If the record value matches with centroid of that cluster, it may remain unchanged. So we replace it with a record's value which is present in the same cluster. Thus we reduce hiding failure.

For simplicity, we have chosen only two clusters. The number of clusters in the overall data is not the matter. For each cluster the process is to be implemented and altered .The same procedure is scalable and can be implemented for any number of clusters.

6. Advantages

(1) Statistical parameters like variance, mean, number of items in a cluster remain the same after anonymization

(2) The procedure implements both K-anonymity and t-closeness for anonymization, hence homogenous attacks are controlled.

(3) The adaptation is done in the supervision of super clusters, hence misclassification errors are avoided.

7. Experimental results

The adult database is taken and clustering and anonymization, calculation of metrics are done using Java. The output is plotted using MATLAB. The scattered plots of original and anonymized databases are shown in the following figures.

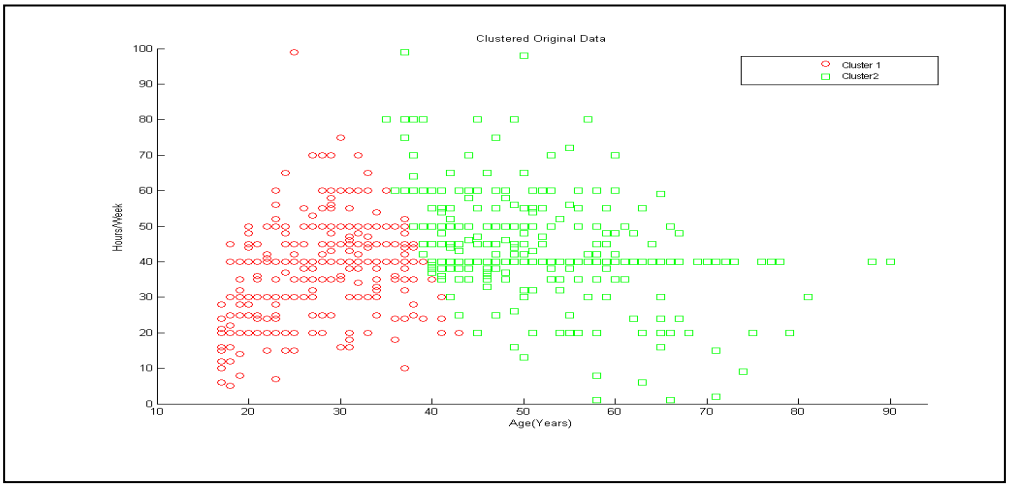

Figure 4: Original Clusters

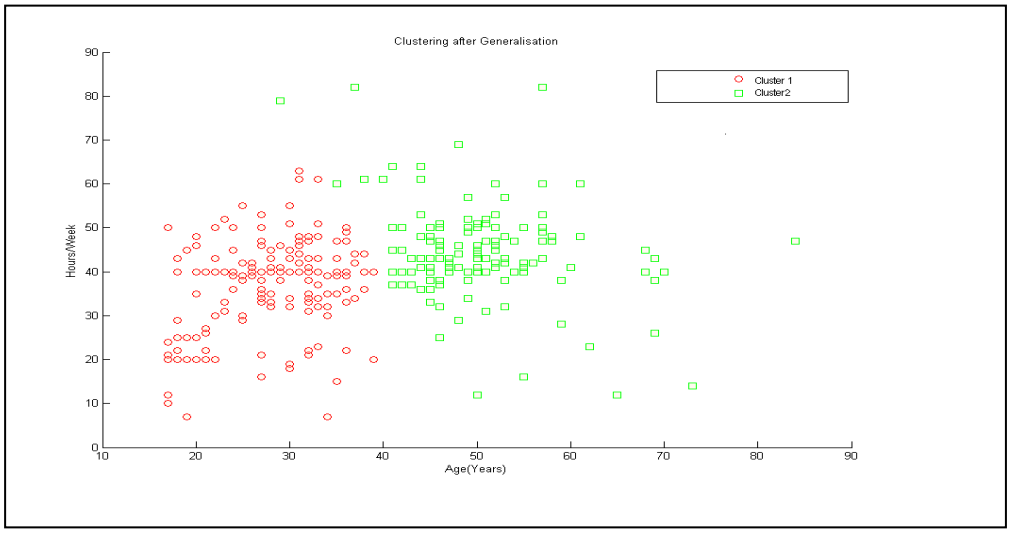

Figure 5: Final Clusters

The output plots show the clustering of objects before and after anonymization. As shown in the plots, the objects are clustered better in anonymized data than the original one. The cluster boundaries are clearer and the objects close to boundary have perturbed by values towards the centroid, hence avoiding the misclassification error. Numbers of records also have the chance of repeated values and hence proving K-Anonymity by itself.

8. Metrics – Explanation and tables

8.1 Evaluation of Anonymization (DA)

This measure is given by Var(X – Y) where X represents a single original attribute and Y the perturbed attribute. Anonymization level can be specified by the metric.

$$DA = Var(X-Y) / Var(X) \quad (7)$$

The measure gives a minimum value, since the perturbation is done within the existing clusters. Also, as the sub – clusters are numbered sequentially, the merging of sub clusters result in lesser variance.

8.2 Evaluation of Clustering (DB)

The efficiency of clustering is measured using Davies–Bouldin index. This index specifies the variation of cluster members within the cluster. During the anonymization process the cluster members are moved towards the centroid of the cluster, so that the boundary overlapping between the clustering is avoided. The scattered records within the cluster become compact during perturbation. This makes the value of this index minimum.

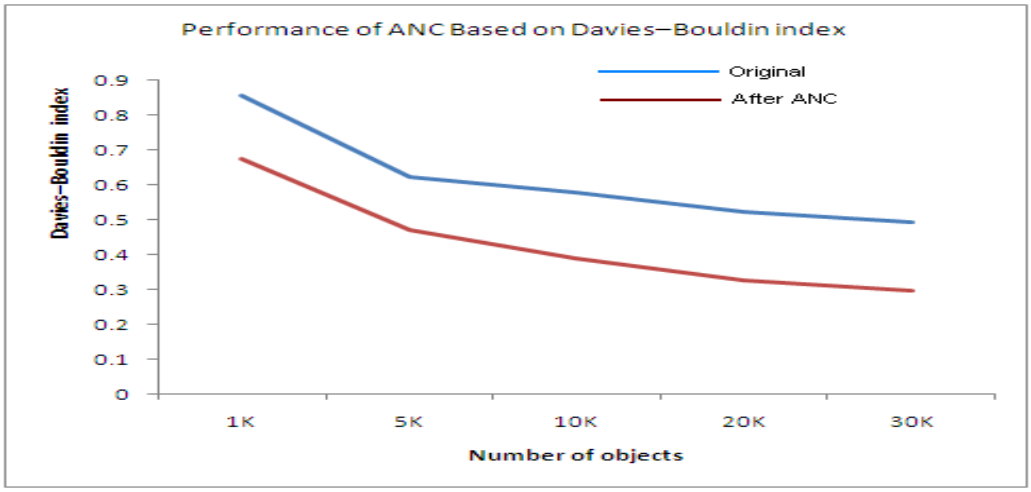

Figure 6 : Performance of ANC based on Davies - Bouldin index

8.3 Mis-Classification Error (ME)

Misclassification error is measured in terms of the percentage of legitimate data points that are not well-classified in the distorted database. A lower misclassification error is desirable as it signifies a smaller change in the object membership of clusters. Ideally, the misclassification error should be 0%. The misclassification error, denoted by ME, is measured as

$$ME = \frac{1}{k}\sum_{i=1}^{n} CM - CM' \quad (8)$$

Where, n – Number of records

K – Number of clusters

CM – Original cluster membership

CM' – Cluster membership after perturbation.

Since the perturbation is based on the parent clustering, the data moves towards the centroid of each of the cluster. Hence it is less likely that the objects are wrongly clustered.

8.4 Computational Costs.

The computation of ANC can be divided into four major steps based on the number of traversal of the whole database. They are

1. Original data clustering
2. Grouping of small sub-clusters
3. Un-grouping of large sub-clusters.
4. Anonymization

Therefore, the records are traversed four times. The proposed data perturbation technique has a complexity of the order O(4 X n) , where n= Number of records . The previous perturbation techniques by [9] and [12] has a complexity of (kX mXn). Hence the performance of ANC is better than the existing ones.

8.5 Hiding Failure

Hiding failure is the portion of sensitive information that is not hidden by the application of a privacy preservation technique. The percentage of sensitive information that is still discovered, after the data has been sanitized gives an estimate of the hiding failure parameter. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure.

$$HF = N/n \quad (9)$$

Where , N = Number of unaltered records which are bound to insecurity

n= Total number of records.

Table 1: Performance of ANC for various metrics.

Metric / Number of Records	DA	DB	ME	HF	CC (ms)
1K	1.637	.6732	0	0	0.2
5K	1.423	.46777	0	0	1
10K	1.385	.38833	0	0	2
20K	1.32	.32642	0	0	4
30K	1.29	.29553	0	0	6

9. Conclusion and Future Work

The experiments, analysis and comparison with existing methods show that, ANC produces a better performance in preserving the privacy of data by not compromising the loss of data. The method has a lower DA values, which shows the amount of distortion of data. This paves the way for a future work to improve this factor. The method can also be implemented on a supervised basis, the perturbation will occur with improved performances.

References

[1] Acar Tamersoy, Grigorios Loukides, Mehmet Ercan Nergiz, Yucel Saygin, and Bradley Malin , "Anonymization of Longitudinal Electronic Medical Records ",IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 3, MAY 2012 .

[2] Aggarwal. C and Yu.P, Privacy-Preserving Data Mining:, Models and Algorithms. Springer, 2008.

[3] Agrawal. R and Srikant. R, "Privacy Preserving Data Mining", Proc. ACM SIGMOD, pp. 439-450, 2000.

[4] Diesburg. S.M and A.-I,Wang.A, "A survey of confidential data storage and deletion methods," *ACM Comput. Surv.*, vol. 43, no. 1, pp. 1–37, 2010.

[5] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data", In Proceedings of the 7th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, pp. 245– 250, New York, USA, 2001.

[6] Fatih Altiparmak, Hakan Ferhatosmanoglu," Incremental Maintenance of Online Summaries over Multiple Streams", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 2, FEBRUARY 2008

[7] Huang.Z, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD, pp. 37-48, 2005.

[8] Karthikeyan. B, Manikandan. G,Vaithiyanathan. V," A FUZZY BASED APPROACH FOR PRIVACY PRESERVING CLUSTERING", Journal of Theoretical and Applied Information Technology,2011,Vol. 32 No.2.

[9] Oliveira. S.R.M and Zaiane.O.R, "A Privacy-Preserving Clustering Approach toward Secure and Effective Data Analysis for Business Collaboration," Computers and Security, vol. 26, no. 1, pp. 81-93, 2007.

[10] Pui Kuen Fong , Weber-Jahnke, J.H.," Privacy Preserving Decision Tree Learning Using Unrealized Data Sets ",Knowledge and Data Engineering,IEEE, Volume:24 Issue:2,2012.

[11] Rajalaxmi.R.R and Natarajan.A.M, "An effective data transformation approach for privacy preserving clustering", Journal of Computer Science 4(4): 320-326, 2008, Science Publications.

[12] Shivaji.S, Dhiraj Ameer M. ,Asif Khan Wajhiulla Khan Ajay Challagalla ,"Privacy Preservation in k-Means Clustering by Cluster Rotation" , IEEE TENCON 2009.

[13] Sweeney, L., Achieving k-anonymity privacy protection using generalization and suppression. 2002.

[14] Vaidya.J and Clifton. C, "Privacy preserving k-means clustering over vertically partitioned data", the 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. ACM Press, 2003.

[15] Yingjie Wu, Zhihui Sun, Xiaodong Wang , "Privacy Preserving k-Anonymity for Re-publication of Incremental Datasets ", 2009 World Congress on Computer Science and Information Engineering.

[16] Yingpeng Sang, Hong Shen, Hui Tian, " Effective Reconstruction of Data Perturbed by Random Projections", IEEE TRANSACTIONS ON COMPUTERS, VOL.61, NO. 1, JANUARY 2012.

Authors Profile

V.Rajalakshmi received M.E from Sathyabama University. She is in the teaching profession for the past 11 years and currently doing research in Privacy preservation in Data Mining. She has published 8 technical papers in various journal / conferences.

Dr.G.S.Anandha Mala received B.E degree from Bharathidhasan University in Computer Science and Engineering in 1992, M.E degree in University of Madras in 2001 and Ph.D degree from Anna University in 2007. Currently she is working as Professor in St.Joseph's College of Engineering, Chennai, India, and heading the department of Computer Science and Engineering. She has published 24 technical papers in various international journal / conferences. She has 18 years of teaching experience on both graduate and post-graduate level. Her area of interest includes Natural Language Processing, Software Engineering, Data Mining, Image Processing and Grid Computing