Isometric Relocation of Data by Sequencing of Sub-Clusters for Privacy Preservation in Data Mining

V.Rajalakshmi [#1], G.S.Anandha Mala[*2]

[#]Research Scholar, Sathyabama University
Chennai, India
[1]rajalakshmi.bala03@gmail.com
[*]Professor & Head,
Dept. of CSE, St.Joseph's College of Engg.
Chennai, India
[2]gs.anandhamala@gmail.com

Abstract—Privacy preservation in data mining is a pioneering research area as the security of increasing amount of data is under risks. Privacy preservation in Data Mining [PPDM] is a delicate task as there is a trade-off between data Anonymization and their utility. Existing PPDM techniques uses Anonymization using randomization, generalization or suppression which reduces the utility of data. They also do not work on the data mining parameters like correlation, centroids etc., This paper provides a solution to handle this trade-off in an efficient way using Isometric relocation. The work uses isometric relocation as it maintains the correlation and data mining results. The methodology is explained with the algorithm and its performance is compared using real-life datasets with existing techniques on various metrics after exhaustive experimentations.

Keywords – Isometric relocation, hierarchical clustering, Privacy Preservation, Anonymization

I. INTRODUCTION

In today's scenario, World Wide Web has increased the number of globally accessible data. Due to this, we are drowning in data but starving for knowledge and privacy. Such data are provided for mining to retrieve non-trivial knowledge for future decision making. As various techniques for revealing non-trivial patterns are explored, the threat towards the data is also increased. When such data are provided as it is for mining it forms a threat for the privacy of the individual. Typical example includes disease of a patient, credit card balance of a customer, purchase details from a departmental store, government weapon details in military, etc., As stated in [17], Anonymization issues also occur in surveying, statistical databases , cryptographic computing, access control, and so on. Hence data need to be modified before they are provided for mining. The crucial part in this process is that modification should not affect the mining result and other statistical parameters about the data. The reason why this modification is accepted is, for mining the exact data is not required, a perfect approximation is sufficient. Therefore, a technique which alters the data without modifying the mining results is termed as PPDM.

Attributes in a database are of three types – unique identifying attributes, sensitive attributes, quasi identifying attributes. When data are given for mining unique identifying attributes like patient ID, credit card number, Employee ID, etc., are removed completely from the database. Sensitive attributes like disease, credit card balance, salary, etc., are the primary concerns for mining and hence they should not be altered. Quasi identifying attributes like age, zipcode, height, married, gender, etc., are also available in a public database like voter's list or known personally by neighbors. These are the values which are altered so that the exact individual of the record is not identified.

The information disclosure is categorized into two types [4], Identity disclosure, specifies which record is associated with which individual in a released table and Attribute disclosure, new information about some individuals is revealed by the released table. In this work, Identity disclosure only is handled i.e., we are trying to hide the association between the individual and the particular record.

PPDM techniques are divided into two categories based on the storage of data, as centralized storage of data or multi party handling of data. The methods which are used for centralized data are of two types – Randomization and Perturbation. In randomization, random numbers are generated with less variance and zero mean are taken. These random numbers are then added with the data in additive perturbation and multiplied in multiplicative perturbation. Cryptographic methods are used for multi party handling of data. Perturbation techniques includes Anonymization, permutation, swapping, slicing, etc., K-Anonymity is one of the widely implemented technique using generalization and suppression. Generalization refers to altering a value with a

less specific but semantically acceptable value, while suppression refers to not releasing a value at all by hiding partially or completely.

PPDM techniques can be classified based on the data mining algorithm for which the method is defined like classification, clustering, decision trees, association rules, etc., They can also be classified based on the privacy preserving technique that is used in the methodology as heuristic, cryptography and reconstructing based. Heuristic methodologies do not consider the relationship of data and hence result in more error during mining. Cryptography based algorithms are generally costlier in terms of execution time and they rely on key values and they have to be reversed for final usage. In this work reconstruction based technique is used which alters the data based on the relationship among them. The other reconstruction based techniques are permutation, swapping, slicing, etc.,

In this work Isometric Relocation of Data by Sequencing of Sub-Clusters for Privacy Preservation in Data Mining [IRSS], concepts of sub-clustering and Isometric rotation using randomization has been used to efficiently anonymize the data, which not only preserves the data but also provides more data utility. The main aim of this algorithm is to provide proper tuning between privacy and utility. This has been done in two steps – varying the size of the sub-clusters and varying the randomness and angles of isometric rotation.

The organization of the paper is as follows- Section 2 explains the various literatures worked in the similar problem and their advantages and disadvantages. Section 3 states the problem definition. Section 4 explains the important concepts used and their explanations. Section 5 provides the detailed algorithm and architecture of the system. Section 6 explores the experimental tools used, analysis of the method on various metrics and graphs with respect to the tuning parameter. Section 7 concludes and mentions some of the future works.

II. RELATED WORK

In [26] Sweeney et al., has started with privacy preservation using k-anonymity. In [21] and [5], Agarwal came up with the technique of perturbation using randomization methods. In [20], Sweeney again introduced a methodology of using a generalization hierarchy for implementing k-anonymity. A new perturbation technique has been suggested in [30] using tree concept. In [25] and [28], Oliviera and et. al. innovated that clustering can be used to group the data for perturbation.[2],[6],[24],[7],[13],[14], and [22] also concentrates on the success of using clustering techniques for implementing k-anonymity or other perturbation techniques. Other than clustering groups of data can be done by nearest neighbor [8], decision tree [27] or bucketization[33].

Oliveira has also identified that isometric transformation based rotation can be used for perturbation. There are various advantages for this method, as it maintains the statistical parameters like centroid, variance, etc., and also best forwards the correlation between the attributes. The only disadvantage of this method being it reversible and allows similar attack. K-anonymity has homogeneity problem and hence improvised as l-diversity[3][17][9] and t-closeness[19].In [24],[25] and [7], isometric transformation is done in clusters, which exhibits the advantages of both of the techniques.

III. PROBLEM DEFINITION

A privacy preserving algorithm in which the data Anonymization is controlled by the relationships among the data, mining parameters like correlation, centroids are not affected in addition to the proper mining output which executes in linear time by not compromising with the privacy is required. The anonymized data should have less distortion compared to the original data.

IV. PRELIMINARIES

There are a few basic definitions need to be known to understand the algorithm and its advantages. Some of them are briefly discussed as follows:

a) Data

Data is information that has been converted and stored in a form that is more comfortable to move or process. Data is usually stored in the form of database arranged as attributes and records. Let T be the database with n number of records and m number of attributes.

b) Quasi Identifier

A set of non-sensitive attributes $\{a_1, \ldots, a_m\}$ of a table is called a *quasi-identifier* if these attributes can be connected with external data to uniquely identify at least one individual in the whole database.

c) Privacy

A database is privacy preserved if there is a minimum probability of associating any transaction with its sensitive attribute.

d) Isometric Rotation

Let T be a transformation in the n-dimensional space, i.e., $F : T_n \rightarrow T_n$ said to be an isometric transformation if it preserves distances satisfying the following constraint: $|F(p)- F(q)| = |p-q|$ for all $p, q \in T_n$.

e) Centroid

It is defined for an attribute as a mean of all the transaction values. All the data are concentrated around this point. After Anonymization the centroid of a cluster is expected to be in-variant. The centroid of cluster k for attribute j is given by,

$$c_{k,j} = \frac{1}{n}\sum_{i=1}^{i=n} r_{i,j} \quad (1)$$

f) Information distortion

Information distortion can be calculated from the difference between the original table and the anonymized table. It can also be calculated as the distribution of data with respect to the centroid. The information distortion can be calculated using the following equations. The dissimilarity of record, i in the jth attribute with respect to centroid cook is given by,

$$diss(r_{ij}, c_{kj}) = [r_{ij} - c_{kj}]^2 \quad (2)$$

The distortion of all records is given by

$$D = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} u_{ik} * diss(r_{ij}, c_{kj}) \quad (3)$$

Where u_{ik} specifies the membership of ith record in kth cluster.

$$\sum_{k=1}^{K} u_{ik} = 1 \quad (4)$$

$$u_{ik} \in \{0,1\} \quad (5)$$

g) Mis-classification error (ME)

The number of records wrongly matched to a different cluster with respect to the total number of records is termed as mis-classification error. The value should be zero for an efficient system. But computationally only for a direct isometric transformation it is zero. Since they are rotated with respect to a point within the cluster the error is minimum compared to a randomization method. As the number of sub-clusters increases, the distance between the centroids decreases resulting in less M_E. There are variations of M_E with respect to simple Isometric transformation as the size of the sub-cluster increases.

h) Amount of Privacy

This is the number of records altered during Anonymization. If the records remain unaltered, then they are said to be unprotected. If the records are homogenous then SIT results in number of unaltered records. As the number of sub-cluster increases, there are more chances for a record to be in equivalence classes and the distance between the centroids become lesser, leading to number of unaltered records. For a perfect system the rate of privacy preservation should be 1 which is impractical, hence an optimal high value is accepted.

i) Fuzzy c-means(FCM) algorithm[1]*:*

A cluster is a group of data points around a center. FCM is a data clustering technique in which each data point belongs to a cluster to some membership value that is specified by a grade. It is mainly used to cluster complex and multi-dimensional data set.

V. ISOMETRIC RELOCATION OF DATA BY SEQUENCING OF SUB-CLUSTERS [IRSS]

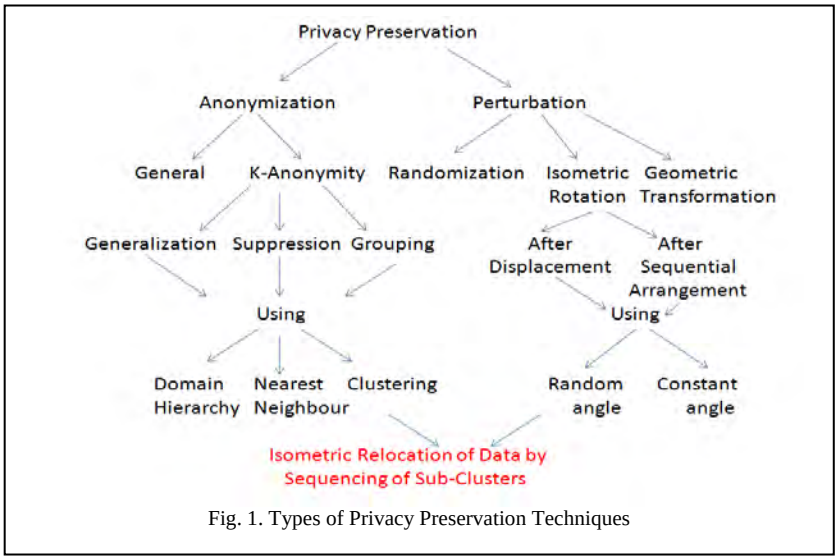

Fig. 1. Types of Privacy Preservation Techniques

In Figure 1, the different types of Anonymization are shown. Figure 3 shows the architecture of the suggested system. In this method, the data are initially grouped into closer ones using clustering. Then each of the clusters is sub-clustered based on the number of records in each cluster. If N is the number of records in each cluster, the number of sub-clusters will be N/3,N/4,N/5,N/6 or N/8. Different numbers were chosen to show that the performance of the algorithm is based on the number of sub-clusters chosen. The sub-clusters are then sequenced based on the Euclidean distance between them. The first sub-cluster will be the closest one with the centroid of the main cluster. For each sub-cluster, the numbers of elements p_i in it are determined. Then p_i numbers of sub-clusters are identified for mapping. Mapping is done towards the centroid of the chosen sub-cluster about a random angle θ, randomly chosen between 10 to 50 radians. Figure 2 shows the isometric rotation of a data with respect to a point.

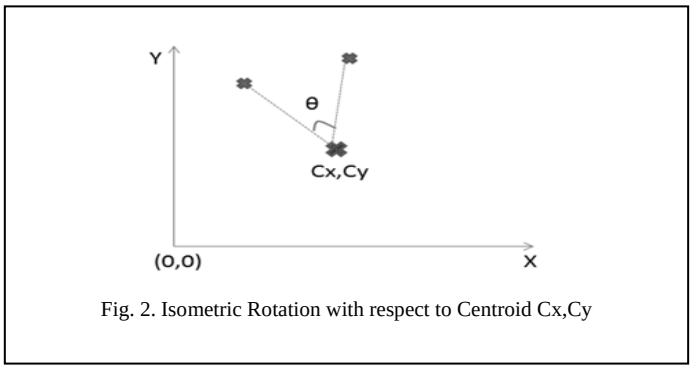

Fig. 2. Isometric Rotation with respect to Centroid Cx,Cy

The method performs well in terms of correlation between the attributes, as it uses isometric rotation of data. Since randomization is not the major part of the procedure, the data are not anonymized more randomly. As the data are first clustered and then anonymized they try to remain within the same cluster. The amount of Anonymization is controlled by two parameters- the size of the sub-clusters and the random angle generated. Table 1 show the mathematical model of the system by specifying the various notations used in the system. Table 2 specifies the different databases, number of attributes used for Anonymization and number of records in them are used for testing the algorithm.

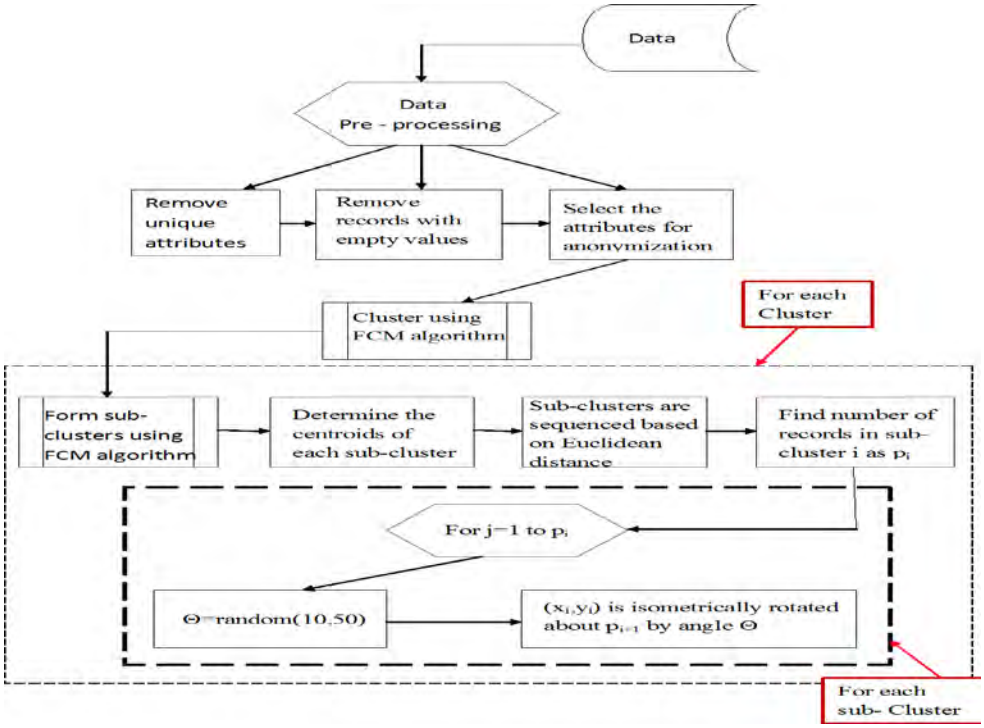

Fig. 3. Architecture of IRSS

Input: Original Database T
Output : Anonymized Database Y

Step1: Data preprocessing

a) Remove unique attributes
b) Remove records with empty values
c) Select the attributes for Anonymization

Step 2: Divide the data into c clusters using FCM algorithm

Step 3: For each cluster

a) Divide the data into K sub-clusters as $S_{1,k}\ldots S_{x,k}\ldots S_{Xk,K}$ using FCM algorithm
b) Let the centroid of each subcluster be $c_1,c_2,\ldots,c_k\ldots c_K$
c) Subclusters are arranged in sequence based on their Euclidean distance
d) Number of elements in sub-cluster i be p
e) For each element $r_{j,k}$
   (i) The angle θ is generated using a randomly between(10,50)
   (ii) $r_{j,k}$ is isometrically rotated with respect to c_{i+1} about an angle θ

Fig. 4. Algorithm of IRSS

TABLE I
Notations Used

Variables used	Explanation
n	Number of records
m	Number of attributes
K	Number of Clusters
C	Centroid
$r_1,r_2,\ldots,r_i\ldots r_n$	Individual record
$a_1,a_2,\ldots,a_j\ldots a_m$	Individual attributes
$1\ldots k\ldots K$	Individual cluster
$c_1,c_2,\ldots,c_k\ldots c_K$	Individual centroid
$X_1,X_2,\ldots X_k\ldots X_K$	Number of sub-clusters in each Cluster
$S_{1,k}\ldots S_{x,k}\ldots S_{Xk,K}$	Individual sub-cluster in cluster k

TABLE III
Database Used

Data base name	No. of records	No. of clusters	No. of attributes used
IRIS dataset	150	3	4
WINES dataset	178	3	4
CREDIT CARD APPROVAL dataset	690	2	4
ADULT dataset	30,162	3	4

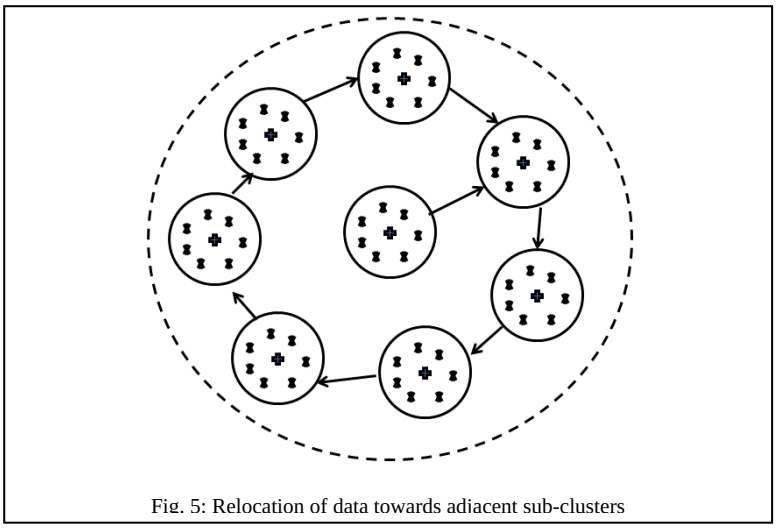

Fig. 5: Relocation of data towards adiacent sub-clusters

The mapping operation is explained in the figure 5.According to this, the sub-clusters are ordered according to their Euclidean distance. For each data in a sub-cluster, in addition to identifying the adjacent sub-cluster find a unique random angle. This will avoid the problem of homogeneity. Though the records belong to the same sub - cluster after anonymization each data will belong to different sub-cluster but within the same main cluster.

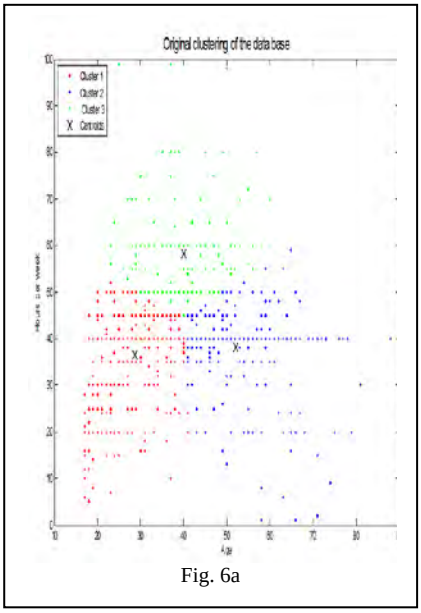

Fig. 6a

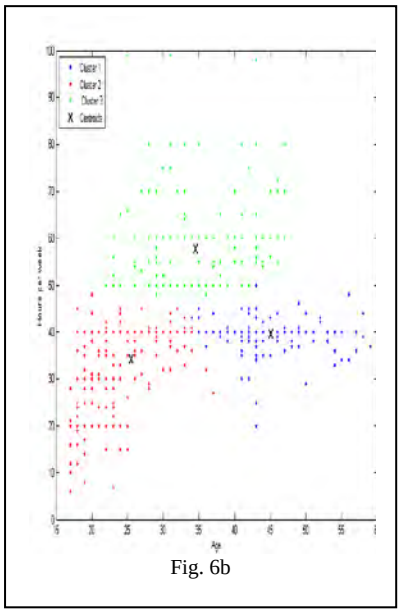

Fig. 6b

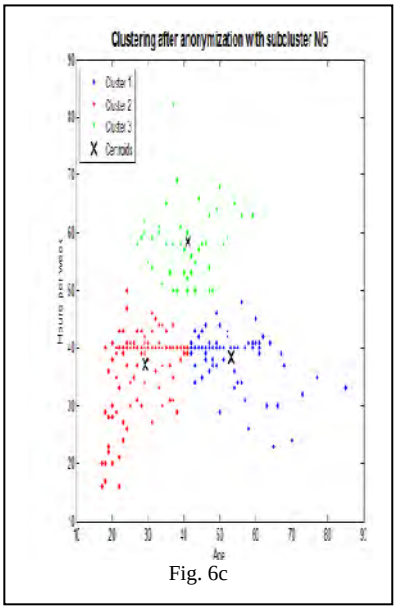

Fig. 6c

Fig. 6a,6b,6c shows the original data clustering, after anonymization with N/3 and N/5 sub-clusters

The methodology is implemented in MATLAB and the results are shown with respect to clustering before and after Anonymization. The performance of the method is also represented in terms of graph for various metrics to compare. Figure 7 shows the performance in terms of information distortion as defined by definition 4.6. If the number of sub-clusters is less as N/8, there will be more data in each sub-cluster, the distance between the centroids will be more and hence the movement of data towards a mapped centroid will be more. This leads to a higher information distortion. Similarly, as the distance between the centroids increases, data tend to move more leading to more classification error. But the amount of privacy preservation is just the opposite. As the size of the sub-cluster is reduced the movements of the data reduces and result in more data not anonymized. Hence, as the number of sub-clusters increases the amount of privacy preservation reduces.

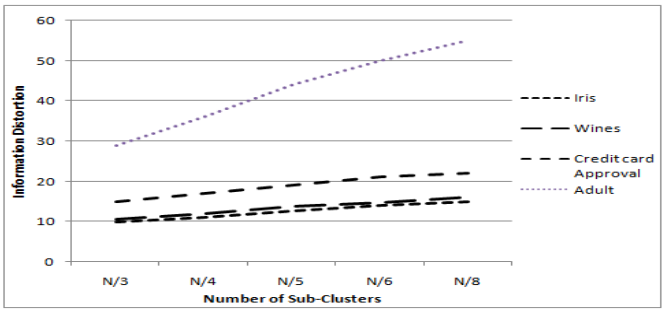

Fig. 7: Performance based on Information Distortion

Figure 8 shows the performance of the algorithm with respect to mis-classification error for various numbers of sub-clusters. The graph shows that as the number of sub-cluster decreases, the number of data in each sub-cluster increases, the distance between the centroids of each sub-cluster increases and hence the mobility of the data increases. This leads to the increase in mis-classification error.

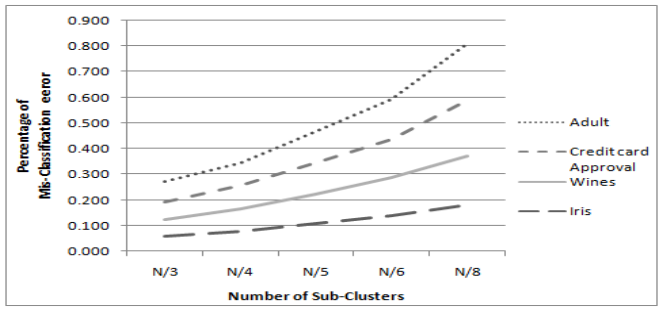

Fig. 8: Performance based on Mis-Classification error

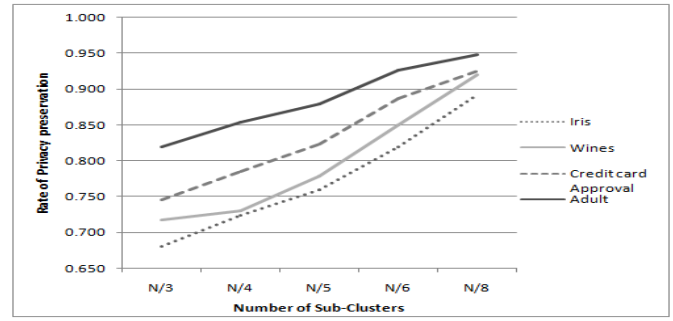

Fig. 9: Performance based on Rate of Privacy Preservation

Figure 9 explains the performance based on the amount of privacy preservation. The graph shows that , as the number of sub-cluster decreases, the mobility of the data increases, leading to more alteration of data and hence more privacy. By figure 6, 7 and 8 it can be concluded that less number of sub-clusters lead to more privacy, but also more information distortion and mis-classification error.

VI. CONCLUSION & FUTURE WORK

Thus the procedure for Isometric Relocation of Data by Sequencing of Sub-Clusters for Privacy Preservation in Data Mining has been explained with all details and the method is tested for various data sets with different sizes of sub-clusters. The method is also evaluated under various metrics for different size of sub-clusters. It can be concluded that the size of sub-clusters highly decides the amount of information distortion, mis-classification error and the rate of privacy preservation. Thus the methodology can provide a tuning between privacy and distortion.

The methodology has been implemented for numeric attributes, which can be extended for alphanumeric attributes and categorical attributes. The methodology is data specific, which can be avoided by building a method using a neural network that can automatically adapt for any kind of data. The methodology can also be improvised for a reduced mis-classification error with more preservation of privacy.

REFERENCES

[1] A.K. Jain, M.N. Murty And P.J. Flynn,"Data Clustering: A Review ",ACM,2000.

[2] Ali _Inan, Selim V. Kaya, Yu cel Saygın , Erkay Savas , Ayc¸a A. Hintoglu,Albert Levi," Privacy preserving clustering on horizontally partitioned data",Data & Knowledge Engineering 63 (2007), 646–666.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam," l -Diversity: Privacy Beyond k-Anonymity ",ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, Publication date: March 2007.

[4] Chai Wah Wu," Privacy preserving data mining with unidirectional interaction", IEEE.

[5] Charu C. Aggarwal and Philip S. Yu," A Condensation Approach to Privacy Preserving Data Mining", Springer-Verlag Berlin Heidelberg 2004,pp 183-199.

[6] Chuang-Cheng Chiu and Chieh-Yuan Tsai,"A k-Anonymity Clustering Method for Effective Data Privacy Preservation",Springer-Verlag Berlin Heidelberg, pp89-99,2007.

[7] Dowon Hong and Abedelaziz Mohaisen," Augmented Rotation-Based Transformation for Privacy-Preserving Data Clustering", ETRI Journal, Volume 32, Number 3, June 2010.

[8] Gabriel Ghinita, Member, Panos Kalnis, and Yufei Tao,"Anonymous Publication of Sensitive Transactional Data ",IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 2, February 2011.

[9] Hongwei Tian and Weining Zhang,"Extending `-Diversity for Better Data Anonymization",NFS grant IIS-0524612,2009.

[10] Ines Buratović, Mario Miličević and Krunoslav Žubrinić,"Effects of Data Anonymization on the Data Mining Results",MIPRO ,2012.

[11] Jaideep Vaidya, Chris Clifton,"Privacy-Preserving Data Mining: Why, How, and When",IEEE SECURITY & PRIVACY,2004.

[12] Jianneng Cao, Barbara Carminati, Elena Ferrari, and Kian-Lee Tan," CASTLE: Continuously Anonymizing Data Streams",IEEE Transactions On Dependable And Secure Computing, Vol. 8, No. 3, May/June 2011.

[13] Ji-Won Byun ,Ashish Kamra ,Elisa Bertino ,Ninghui Li,"Efficient K-Anonymity Using Clustering Technique",Cerias Tech Report 2006-10.

[14] Kun Guo a, Qishan Zhang," Fast clustering-based anonymization approaches with time constraints for data streams",Knowledge-Based Systems 46 (2013) 95–108.

[15] Latanya Sweeney,"Achieving K-Anonymity Privacy Protection Using Generalization And Suppression",International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571- 588.

[16] M.E. Nergiz, M. Atzori, and C. Clifton, “Hiding the Presence of Individuals from Shared Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 665-676, 2007.

[17] Machanavajjhala A, Gehrke J, Kifer D, l-diversity: Privacy beyond kanonymity,Proc of the International Conference on Data Engineering,Atlanta, GA, USA, 2006, pp. 24.

[18] N.R. Adam and J.C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparative Study,” ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[19] Ninghui Li ,Tiancheng Li,Suresh Venkatasubramanian," t-Closeness: Privacy Beyond k-Anonymity and -Diversity",2007, IEEE.

[20] Pui K. Fong and Jens H. Weber-Jahnke,"Privacy Preserving Decision Tree Learning Using Unrealized Data Sets ",IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 2, February 2012.

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

[22] R. Vidya Banu a, N. Nagaveni," Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario",Information Sciences 232 (2013) 437–448.

[23] REN Xiangmin,YANG Jing," Research on privacy protection based on K-anonymity",2010,IEEE

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

[25] S.R.M. Oliveira and O.R. Zaïane, “Achieving Privacy Preservation When Sharing Data for Clustering,” Proc. SDM, 2004, pp. 67-82.

[26] Samarati P, Sweeney L, Generalizing data to provide anonymity when disclosing information, Proceedings of the Seventeenth ACM SIGACT SIGMOD-SIGART Symposium on Principles of Database Systems,PODS, Seattle,WA, USA, 1998, p.188.

[27] Slava Kisilevich, Lior Rokach, Yuval Elovici, Member, IEEE, and Bracha Shapira," Efficient Multidimensional Suppression for K-Anonymity",IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2010.

[28] Stanley R. M. Oliveira, Osmar R. Zaiane," Data Perturbation by Rotation for Privacy-Preserving Clustering",Technical Report TR 04-17 August 2004.

[29] V. Iyengar, “Transforming Data to Satisfy Privacy Constraints,”,Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 279-288, 2002.

[30] Xiao-Bai Li and Sumit Sarkar," A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining",IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 9, September 2006.

[31] Xiaoxun Sun, Min Li , Hua Wang," A family of enhanced (L, α)-diversity models for privacy preserving data publishing",Future Generation Computer Systems 27 (2011) 348–356.

[32] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang,"Enabling Multilevel Trust in Privacy B26Preserving Data Mining",IEEE Transactions On Knowledge And Data Engineering, Vol. 24,2012.

[33] Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, and Donghui Zhang,"ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication ", IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 7, July 2009.