A Hybrid Approach in Privacy Preserving Data Mining

Miss.Anjana Patel , Ass.Prof Khyati Patel

Computer Department, Sliver oak collage of Engineering and Technology , Ahmedabad, Gujrat, India

ABSTRACT

Information sharing between two associations is basic in numerous application areas like business planning or marketing. When information are to be shared between gatherings, there could be a few sensitive data which should not be disclosed to the next parties .Also medical records are more delicate so, privacy protection is taken more seriously. As per requirement by the Health Insurance Portability and Accountability Act (HIPAA), it is necessary to protect the privacy of patients and ensure the security of the medical data. we propose a method called Hybrid approach for privacy preserving. First we randomizing the original data. Then we apply geometric data Transformation on randomized or modified data. Modify k-means clustering algorithm is used to check the correctness. This technique protect private data with better accuracy, also it can reconstruct original data and provide data with no information loss, makes usability of data.

Keywords: *Privacy preserving; Data Mining; Sensitive Data; Randomization; Geometric data Transformation.*

I INTRODUCTION

Huge volume of detailed private data are regularly collected and analyzed by applications using data mining. Such data include shopping habits, criminal records, medical history, credit records etc. On one hand, such data is an important benefit to business organizations and governments both for decision making processes and for providing social benefits, such as medical research, crime downsizing, national security, etc [1]. On the other hand, the analysis of such data opens new threats to privacy and autonomy of the individual if not done properly. The ultimatum to privacy becomes real since data mining techniques are able to derive very sensitive knowledge from unclassified data that is not even known to database owners.

Clustering is a widely used data mining technique in many applications such as customer behaviour analysis, targeted marketing, and many others. Achieving privacy preservation when sharing data for clustering is a challenging problem. To address this problem, the system must not only meet privacy requirements of the data owners but also guarantee valid clustering results. As found in [2], two major constraints in this context are:

- Distinguishing of patients with similar disease(s) without revealing the values of the confidential attributes associated with them; and
- Two organizations having dataset with different attributes for a common set of individuals can share the cluster results without learning anything about the attribute values of each other;

The first constraint of PPC (privacy-preserving clustering) can be represented as: (i) PPC over centralized data and (ii) PPC over horizontally partitioned data , however, the second constraint can be represented as (iii) PPC over vertically partitioned data. The problem of PPC over vertically and horizontally partitioned data has been addressed in the literature [3,4], while the problem of PPC over centralized data has not been significantly address. In this paper, focus is given mainly on PPC over centralized data. However, the proposed method with little modification can also be applied for the other constraint.

Our survey reveals that very less work has been done on the problem of PPC over centralized data. A significant work can be found in [5]. However, a key finding of our study is that adding noise to data would meet privacy

requirements, but it may lead to compromise of the clustering analysis. The main problem is that by distorting the data, many data points would move from one cluster to another jeopardizing the notion of similarity between points in the global space. Consequently, this introduces the problem of misclassification. This paper attempts to address this issue by introducing a Hybrid Data Perturbation (HDP) method which can be found to be significant in view of the following points:

- more secure
- accuracy is maintained
- there are scopes for further improvement in security without losing accuracy measure
- easy to implement
- independent of any distance based clustering technique

The rest of the paper is organized as follows. Some of the relevant works are reported in section 2. Background of the work is reported in section 3. Section 4 presents the methods. In section 5, the method is evaluated in terms of accuracy & security. Finally, in section 6 conclusions is reported.

II. RELATED WORK

Based on our survey related to PPDM (Privacy Preserving Data Mining), it has been observed that most of the existing works are aimed at providing solutions restricted basically to data partition [3,4], randomization and data distortion. In this section we report some of these in brief.

A. Savita Lohiya & Lata Ragha Privacy Preserving in Data Mining Using Hybrid Approach[2].

The proposed Hybrid approach employs randomization and K-anonymity method. By using Randomization method attacker cannot identify a design of data. K-anonymity method has shortcoming of homogeneity and background attack.In the proposed method as we combined K-anonymity with randomization method it is difficult for attacker to identify homogeneity and background attack. Also it protects private data with better accuracy and gives no loss of data which makes usability of data, and also data can be reconstructed.

B. Vaidya & Clifton's Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data [3]

It addresses the clustering of vertically partitioned data. In the vertical partitioning the attributes of the same objects are split across the partitions. Here, it introduces a solution based on security multi-part computation. Specifically, the authors proposed a method for k-means clustering when discrete sites contain different attributes for a common set of entities. In this solution, each site learns the cluster of each person, but learns nothing about the attributes at other sites. This work ensures reasonable privacy while restricting communication cost.

C. Meregu and Ghosh's Privacy-Preserving Distributed Clustering Using Generative Models [4]

A new model is proposed based on generative models to address privacy preserving distributed clustering. In this approach, rather than sharing parts of the original data or perturbed data, the parameters of suitable generative models are construct at each local site. Then such parameters are transfer to a central location. The best representative of all data is a certain "mean" model. It was empirically shown that such a model can be approximated by creating artificial samples from the underlying distributions using Markov Chain Monte Carlo techniques This methodology accomplishes excellent distributed clustering with allowable privacy loss and low communication cost.

D. Oliveira and Zaiane's Privacy Preserving Clustering By Data Transformation [5]

Here, the feasibility of achieving PPC through geometric data transformation is studied. This inspection revealed that geometric data transformations, such as translation, scaling, and simple rotation are infeasible for privacy preserving clustering if we do not consider the standardization of the data before transformation. The cause is that the data transformed through these methods would change the equality between data points. As a result, the data shared for clustering would be worthless. This work also revealed that the abnormality methods adopted to successfully stable privacy and security in statistical databases are constrained when the irritated qualities are considered as a vector in the n-dimensional space. Such methods would exacerbate the issue of misclassification. A promising direction of the work in [5] was that PPC through data transformation should be to some extent practical by isometric transformations, i.e., transformations that preserve distances of objects in the process of moving them in the Euclidean space.

E. M Kalita,D K Bhattacharyya , M Dutta Privacy preserving clustering –A hybrid Approach[11]

This paper presents a privacy preserving clustering technique using hybrid approach. The technique mainly utilize a combination of isometric transformations i.e. translation, rotation and reflection transformations along with a secure random function in sequence to provide secrecy of user-specified attributes without misplace, accuracy in results. The proposed method was tested and evaluated in terms of several synthetic as well as real-life data and the performance has been establish satisfactory in comparison to its other counterparts. In this paper, focus is given mainly on PPC over centralized data. However, the proposed method with little modification can also be applied for the other constraint.A key finding of our study is that adding noise to data would meet privacy requirements, but it may lead to compromise of the clustering analysis. The main problem is that by distorting the data, many data points would transfer from one cluster to another jeopardizing the notion of similarity between points in the global space. Consequently, this introduces the difficulty of misclassification.

F. G. Manikandan, N. Sairam, C. Saranya and S. Jayashree A Hybrid Privacy Preserving Approach in Data Mining[8]
In this paper we put forward a hybrid approach for achieving privacy while the mining procedure. The first step is to sanitize the original data using a geometrical data transformation. In the second step this sanitized data is normalized using a min-max normalization approach before publishing. It can also be decide, that the elements are scattered to a larger scale in modified data from a confined range in original data. Thus, without knowing the range of the original data, the actual sensitive data can never be identified and the privacy is preserved.

G. Discussions
Based on our limited survey following observations are made:

- Several methods have been proposed in the current decade to address the privacy preserving clustering issues. However, those methods based on isometric concept have been found more suitable for the purpose.
- Isometric transformations, i.e. Basically the three fundamental transformations- translation, rotation and reflection play a key role in providing a better tradeoff between the privacy & precision.
- Translation, rotation, and reflection are the three fundamental isometric transformations which keep the distances between the corresponding objects invariant after transformation.
- Perturbation using translation and rotation are already implemented. However, in both these cases the randomization function plays a very crucial rule. If the function is not properly chosen it may lead to humiliation of cluster quality.
- A Hybrid method using the variants of these three basic transformations along with an appropriate randomization function seems to be a better choice.

III. BACKGROUND OF THE WORK

A. Randomization Technique:-

Randomization was initially used in the context of survey which have privacy con-cerns . It was introduced to preserve privacy data mining by Agrawal and Srikat .In randomization, noise was added to the data so that the individual values of the records cannot be recovered. However the probability distribution of the aggregate data can be recovered and be used for data mining. Representative randomization methods include the additive-noise-based perturbation, the projection-based perturbation and the Randomized Response. The additive-noise-based perturbation can be described as follows:

$$Y = X + E$$

We use X to denote the original data, E to denote the additive noise and Y to represent the perturbed data. Let Xj be the j-th column of the original micro data table corresponding to a sensitive attribute and suppose that there are N tuples. Each value xij (i = 1; : : : ;N), is replaced by:

$$yij = xij + ²ij$$

However, this kind of randomization is not secure under some attacks.

B. Concept of Geometric Data Transformations

Transformations which leaves the metric properties of the area unaltered are called isometric. Under these transformations the space is not stretched or warped so that the distances between any pair of points remain unchanged upon transformation. Formally, an geometric transformation is defined as follows:

geometric transformation include: (1) Translations, which shift points a constant separation in parallel directions; (2) Rotations, which have a center a such that $|T(p) - a| = |p - a|$ for all p; and (3) Shearing, Method Such that each confidential data will takes only multiplicative noise For the sake of simplicity, such a transformation is done in a 2D discrete space.

B 1.Translation Based Perturbation:-

In this method the noise word applied to each confidential attribute is constant and can be either positive or negative. The set of operations takes only the value {Add} compatible, to an additive noise applied to each confidential attribute.

Translated value = Original value + Noise.

B 2. Rotation Based Perturbation:-

In this method a rotation matrix is used to rotate two attributes at a time. For the purpose of straightforwardness a 2D rotation matrix is considered. The rotation of a point by an angle θ in a 2D discrete space can be seen as a matrix representation V_ = Ro(θ)×V, where V is the column vector containing the original coordinates, and V_ is a column vector whose coordinates are rotated coordinates and Ro(θ) is a 2×2 rotation matrix,

$$R0(\theta) = \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix}$$

B.3 Shearing Based Perturbation:-

A noise is added to each confidential data in the Shearing Data Perturbation Method Such that each confidential data will take only multiplicative noise. Sheared value = Original value + (Noise * Original value)

B.4 Min-Max Normalization:-

Min-max normalization performs a linear transformation on the original data. Each characteristic is normalized by scaling its values so that they fall within a small specific range, for example, 0.0 and 1.0. Min-max normalization maps a value V of an attribute A to V' as follows:

$$V' = \frac{V - min_A}{max_A - min_A} x\,(new\_max_A - new\_min_A) + new\_min_A$$

where v' is the new value in the required range. The advantage of Min-Max normalization is that it preserves the relationships among the data values. Thus, without knowing the range of the original data, the actual sensitive data can never be identified and the privacy is preserved.

IV. PROPOSED METHOD

Through the study of literature survey has identified different privacy preserving techniques still there is a need to find the efficient technique to preserve privacy in large database. The proposed Hybrid algorithm has two main advantages. The proposed method protects private data with no loss of data which makes usability of data, and also data is reconstructed.

A. Hybrid Algorithm

1) Inputs: Original training dataset T, Transition probability matrix P, Mapping Matrix M with size 1*j between T and P

2)Output:Conversed training dataset D, Derived table.

3) Method:

a) Select the attribute from table T.

b)Generate probability matrix P randomly with size j*j.

c)Generate mapping matrix M randomly

d)assign each P (P1, P2… Pj) to the column of T (T1, T2… Tj) randomly by M.

e) Rearrange element of T with respect to highest value of P location. If P location is already employ go for next highest location, if value of the P of two or more location is same then choose the left hand side value.

f) Recombine T matrix.

g) Re-substitute in table.

h)Apply Geometric Data Transformation namely Translation, Rotation and Shearing.

i) Apply Normalization on the Transformed data

j) Stop.

The proposed algorithm first selects attribute or quasi Identifier from table T. After that it generates probability matrix P and mapping matrix M randomly. Element of T are rearranged with respect to highest value of P location. If P location is already used, then it will go for next highest location, if value of the P of two or more location is same then it will choose the left hand side value. All values are re substituted in table T. After that numeric values

are Translate, Rotate and shear using Geometric data Transformation and after that Min-Max Normalization is used to normalized the sanitized data.

K-means clustering have a some drawback like center update and find a distance between every value so the computation overhead is become high and performance is decrease .so we can some modification in k-means clustering .we can fix the center of the cluster using the partition of the data using the minimum and maximum value of the data. Then we partition the dataset into equal size and fix the center for one cluster so whenever new data is arrived center updating is not necessary. Using modify k-means we can achieve a reliability between data.

Procedure:

Step 1:User request's data from the Coordinator
Step 2:Coordinator identifies the sensitive data in the data set
Step3:-Sensitive data is randomized using Randomization.
Step4:Confidential Data is modified using Translation and Rotation and Shearing
Step5:Data is then normalized using Min-Max Normalization Process
Step 6: The sanitized data is given to the client
Step 7:Client Uses Modify K-means algorithm for clustering process

V. RESULT AND ANALYSIS

Table 1:-clustering result using modify k-means

K=2	Cluster1	Cluster2
Original Data	{21,33,22,18,27}	{45,46,58,39}
Geometric Data After Transformation	{73,60}	{77,128,130,137,139}
Transform Data After Normalization	{10,45,37,49}	{85,87,92,94,100}
After Shearing and Normalization	{28,37,47,48}	{69,69,73,71,83}

Table 2: Comparison

Original Data	Randomization	Noise=7 Transformed data	Transformed data After Normalization	noise=pi/5 Rotation Data	noise=1.468 Shering Data	Normalized data After Shering with Min-Max
21	33	41	10	10	10	28
33	45	49	16	15	38	37
45	18	75	34	30	75	47
22	21	78	36	32	79	48
46	58	128	70	60	149	69
58	46	129	71	61	151	69
18	39	137	76	66	163	73
27	27	130	72	63	156	71
39	22	172	100	87	215	88

In this paper, we have used Randomization and geometrical transformations namely translation and Rotation and shearing followed by min-max normalization to achieve privacy and accuracy and Secrecy of the data during data mining and modify K-means clustering is used to check the Correctness. Here the computations for Randomization Technique, Geometric transformation and min-max normalization of sample data, Modify k-means clustering and effectiveness calculations are carried out in Matlab .We have also tested the efficiency of our approach on a real-time dataset,' adult-dataset' from UCI data repository[15].

The following snapshots are based on a sample data set containing 9 elements, which are 21,33,45, 22,46,58,30,18,27,39. The clustering of original, Transformed,Transfrom normalized data and Shearing After Normalization for 2-clusters is given in the Figures 1,2,3 and 4 respectively.
Tables 1 describe the clustering of data before and after min-max normalization for 2-clustering. In Table 2 Comparison of the result is shown.

```
After applying Modified K-means Clustering on Original Data - Cluster Values are:
------------------------------
Cluster1:
No. of Values in Cluster 1 in Two Partition is 4
21
22
18
27
Cluster2:
No. of Values in Cluster 2 in Two Partition is 5
33
45
46
58
39
------------------------------
```

Fig 1: clustering on original data

```
After Data Transformation Cluster Values are:
------------------------------
Cluster1:
No. of Values in Cluster 1 in Two Partition is 4
41
49
75
78
Cluster2:
No. of Values in Cluster 2 in Two Partition is 5
128
129
137
130
172
```

Fig 2:-Clustering After Data Transformation.

```
After Normalization using Min-Max Algo - Cluster Values are:
------------------------------
Cluster1:
No. of Values in Cluster 1 in Two Partition is 4
10
16
34
36
Cluster2:
No. of Values in Cluster 2 in Two Partition is 5
70
71
76
72
100
```

Fig 3:-Clustering After Normalization.

```
------------------------------
After Shearing & Normalization - Cluster Values are:
------------------------------
Cluster1:
No. of Values in Cluster 1 in Two Partition is 4
28
37
47
48
Cluster2:
No. of Values in Cluster 2 in Two Partition is 5
69
69
73
71
88
------------------------------
```

Fig 4:-Clustering After Shearing and Normalization using min-max Normalization.

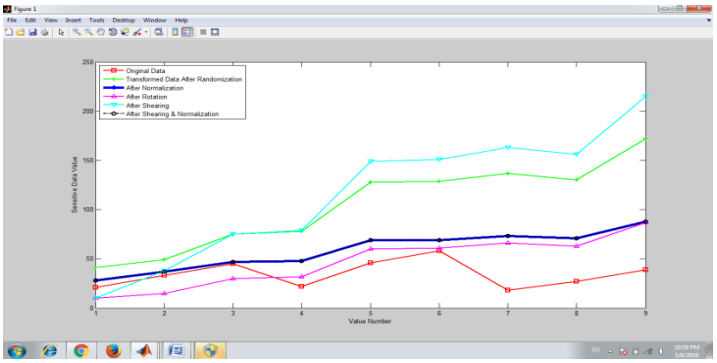

Fig 5: Comparision Graph

As per the resulting Graph it is clear that we can achieving higher privacy and accuracy of the data as well as security of the data is also achieved using the more layer of perturbation of the noise so Combination of Randomization and Geometric Data Transformation gives higher privacy and better accuracy .Using Modify k-means clustering technique center is not updated every time when new data element is arrived. so we can get a better performance and reduce the computation overhead. Data Element in every cluster is not similar so Security of the data is Obtain.

VI. CONCLUSION

An ultimate goal for all data perturbation algorithm is to optimize the data transformation process by maximizing both data privacy and data utility achieving. Proposed approach focused on data perturbation by Geometric transformation and noise addition to preserve privacy of sensitive attributes. We are going to combined two technique of data mining is randomization and geometric transformation for providing security of the sensitive data. We consider single attribute as sensitive attributes or dependent attributes and rest are as non sensitive attributes. We will evaluate the experiment result in terms of correctly distribute the data in clustering and higher privacy, accuracy and the no information loss with the security of the data. Result will show fairly good level of privacy has been achieved with reasonable accuracy and security in almost all tested cases.

REFERENCES

1. Jiawei Han and Micheline Kamber, 2006. Data Mining-Concepts and Techniques, 2nd. Edition. San Francisco: Morgan Kaufmann Publishers.
2. Savita lohiya,Lata Ragha,2012."Privacy Preserving in Data Mining Using Hybrid Approach",IEEE
3. J. Vaidya and C. Clifton. Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data. In *Proc. of the 9th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining*, pages 206–215, Washington, DC, USA, August 2003.
4. S. Meregu and J. Ghosh. Privacy-Preserving Distributed Clustering Using Generative Models. In Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03), pages 211–218, Melbourne, Florida, USA, November 2003.
5. S. R. M. Oliveira and O. R. Zaiane. Privacy Preserving Clustering By Data Transformation.In Proc. of the 18th Brazilian Symposium on Databases, pages 304–318, Manaus, Brazil, October 2003.
6. Mr.S.Chidambaram, Research Scholar/NEC, Dr.K.G Srinivasagan, Professor/CSE/NEC," A Combined Random Noise Perturbation Approach for Multi Level Privacy Preservation in Data Mining", International Conference on Recent Trends in Information Technology,IEEE 2014.
7. Dr.Ruvan Kumara Abeysekara and Prof. Weishi Zhang," Hybrid Framework for Privacy Preserving Data Sharing", International Conference on Advances in ICT for Emerging Regions (ICTer): 198 – 206,IEEE 2013.
8. G. Manikandan, N. Sairam, C. Saranya and S. Jayashree," A Hybrid Privacy Preserving Approach in Data Mining", Middle-East Journal of Scientific Research 15 (4): 581-585, 2013
9 Manikandan, G., N. Sairam, R. Sudhan and Vaishnavi, . 2012. "Shearing Based Data Transformation Approach for Privacy Preserving Clustering" , In. Proceedings of 3rd IEEE International Conference on Computing, Communication and Networking . Technologies, ICCCNT.

10. Liming Li, Qishan Zhang," A Privacy Preserving Clustering Technique Using Hybrid Data Transformation Method", International Conference on Grey Systems and Intelligent Services, November 10-12, 2009, Nanjing, China,IEEE 2009.
11. M Kalita,D K Bhattacharyya and M Dutta,2008,Privacy Preserving Clustering- A Hybrid Approach,ln Proc. ADCOM,IEEE
12. Rajalaxmi, R.R. and A.M.Natarajan, 2008. "An Effective Data Transformation Approach for Privacy Preserving Clustering", Journal of Computer Science, 4(4): 320-326.
13. Karthikeyan, B.G.Manikandan and V.Vaithiyanathan, 2011. "A Fuzzy Based Approach for Privacy Preserving Clustering", Journal of Theoretical and applied information Technology, 32(2): 118-122.
14. Ming-chuan hung, jungpin wu+, jin-hua chang and don-lin yang," An Efficient k-Means Clustering Algorithm Using Simple Partitioning*"
15. UCI Data Repository: http://archive.ics .uci.edu/ml/datasets.html.