

## A Hybrid Privacy Preserving Approach in Data Mining

*G. Manikandan, N. Sairam, C. Saranya and S. Jayashree*

School Of Computing, SASTRA University, Thanjavur, India - 613402

**Abstract:** Data mining algorithms extracts the unknown interesting patterns from large collection of data set. Some clandestine or secret information may be exposed as part of the data mining process. In this paper we put forward a hybrid approach for achieving privacy during the mining procedure. The first step is to sanitize the original data using a geometrical data transformation. In the second stage this sanitized data is normalized using a min-max normalization approach before publishing. For verifying the experimental results we have used k-means clustering algorithm and it is evident from our results that this hybrid approach preserves privacy and also ensures accuracy.

**Key words:** Accuracy • Clustering • K-Means • Min-Max normalization • Privacy

### INTRODUCTION

A quality check must be carried out before a new algorithm or a product is released. Exhaustive testing is performed in order to check the quality of the product and the efficiency of an algorithm. For this testing procedure, a synthetic data set can be used. This synthetic data may not reflect the real time environment which may lead to the failure of the algorithm in the real time [1]. To overcome this, the algorithm is tested with real time data set which can be retrieved from the data providers/vendors. In order to maintain the integrity, the data vendors may sanitize the original data and then transmit them for testing. The central theme behind data sanitization is to maintain privacy without compromising accuracy. In the data mining domain, while performing mining operation, data sanitization is done to ensure privacy. A number of privacy preserving algorithms have been proposed and are in use today. In this paper, we propose a new hybrid method for achieving preserving data during data mining. This approach is a two step process. In the first step the sensitive data is modified using geometrical transformations namely Shearing and Translation. The Original data is shown in Table 1(a) and the Transformed data is shown From the Table 1(b), a user can easily infer that the data is not the original one but it is a sanitized data, based on the value of age attribute in the fifth record. The user should be concealed from the fact that the data is modified. To overcome this we use a normalization process in the second step. The Output of this step is shown in Table 2.

Table 1(a): Original Data

Id	Name	Age	Gender
1	Anu	10	F
2	Saran	20	F
3	Hari	25	M
4	Jaya	30	F
5	Adhav	50	M

Table 1(b): Transformed Data

Id	Name	Age	Gender
1	Anu	24	F
2	Saran	44	F
3	Hari	54	M
4	Jaya	64	F
5	Adhav	104	M

Table 2: Min-Max Normalized Data

Id	Name	Age	Gender
1	Anu	10	F
2	Saran	30	F
3	Hari	40	M
4	Jaya	50	F
5	Adhav	90	M

Various types of normalization like Min-Max, Z-score and decimal scaling can be used. To verify the correctness, we use K-means clustering algorithm to the original and sanitized data.

The rest of the paper is organized as follows: Section II provides an overview of literature works carried out in Privacy techniques; Section III elaborates the implementation of hybrid approach in our work. Simulation snapshots and the outcome of

experimental results are discussed in Section IV and finally in Section V, we arrive to an overall conclusion from our work.

**Literature Review:** Uses an Effective Data Transformation Approach for Privacy Preserving Clustering Categorical Data by using a set of hybrid geometrical data transformations such as HDTTR and HDSTR [2].

S-shaped membership function for data sanitization and Fuzzy Logic can be used to achieve Privacy. These methods were demonstrated in [3].

Proposed a shearing based composite data transformation approach for achieving privacy[4]. A set of composite transformations like CDTTSh, CDTShT, CDTShS, CDTSSh were used for perturbing the data. Modified data depends on the noise value.

New Noise addition methods based on pre-mining to protect health care privacy was proposed in [5].

Uses Isometric Transformation [6] for achieving privacy by selecting a pair of attributes and then distorts them with select angles.

**Proposed System:** Usually in data perturbation, the data is sanitized using a single noise value. But the data perturbation in our approach is based on the use of two noise values. Here we use general line equation as a model for data sanitization. The line equation is given by  $y = mx + c$ . Where,  $x$  is the original data,  $m$  and  $c$  are noise values and  $y$  is the perturbed data. In this section, we give a brief note on the following techniques namely Shearing, Translation and Min-Max normalization.

**Translation and Shearing:** Translation is defined as repositioning of an object along a straight line path, from one coordinate location to another. Shearing is also a type of translation that changes the shape of an object causing it to slide over by crushing its internal layers [7].

**Min-Max Normalization:** Min-max normalization performs a linear transformation on the original data. For mapping a value,  $v$  of an attribute  $A$  from range  $[\min_A, \max_A]$  to a new range  $[\text{new\_min}_A, \text{new\_max}_A]$ , the computation is given by

$$\frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

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. Table 1(a) is the sample data set used for experiment.

Table 2 is the corresponding normalized values for the Transformed data after applying min-max normalization. Figure 1 shows the flow diagram for the proposed system and the steps involved in our approach can be summarized in the form of a procedure as shown below

#### Procedure:

- Step 1: User request's data from the Coordinator.
- Step 2: Coordinator identifies the sensitive data in the data set.
- Step 3: Confidential Data is modified using Shearing and Translation
- Step 4: Data is then normalized using Min-Max Normalization Process
- Step 5: The sanitized data is given to the client.
- Step 6: Client Uses K-means algorithm for clustering process.

**Simulation and Results:** In this paper, we have used geometrical transformations namely translation and shearing followed by min-max normalization to achieve privacy and accuracy during data mining and K-means clustering is used to check the accuracy. Here the computations for Geometric transformation and min-max normalization of sample data, k-means clustering and effectiveness calculations are carried out in Java Developer kit with a synthetic data set of varying sizes like 1k, 2k etc. We have also tested the efficiency of our approach on a real-time dataset, 'adult-dataset' from UCI data repository [8].

The following snapshots are based on a sample data set containing 9 elements, which are 2, 4, 10, 25, 11, 3, 30, 20, 12. The clustering of original, Transformed and the normalized data for 2-clusters is given in the Figures 2, 3 and 4 respectively. Similarly, those for 3-clusters are provided in the Figures (5) and (6).

Tables 3 and 4 describe the clustering of data before and after min-max normalization for 2-clustering and 3-clustering respectively. Figure 8 & 9 clearly illustrates that the data is clustered exactly the same in all the three forms namely Original data, Transformed data and Normalized data without any misclassification errors. This is evident from the below figures that the number of elements in each clusters never changed and remained the same.

Table 3: Results for 2-clusters

K=2	Cluster 1	Cluster 2
Original Data	{2,4,10,12,3,11}	{20,30,25}
Geometric Data Before Normalization	{8,12,24,28,10,26}	{44,64,54}
Geometric Data After Normalization	{10,15,32,38,12,35}	{60,90,75}

Table 4: Results for 3-clusters\

K=3	Cluster 1	Cluster 2	Cluster 3
Original Data	{2,4,3}	{10,12,11}	{20,30,25}
Geometric Data Before Normalization	{8,12,10}	{24,28,26}	{44,64,54}
Geometric Data After Normalization	{10,15,12}	{32,38,35}	{61,90,75}

Table 5: Comparison

Original Data	Transformed data (Noise=2,4)	Normalized data with Min-Max
2	8	10
4	12	15
10	24	32
12	28	38
3	10	12
11	26	35
20	44	61
30	64	90
25	54	75

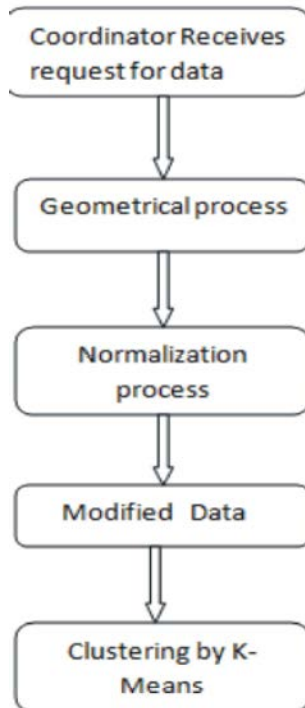


Fig. 1: Flow Diagram for proposed system

```

Clustered values are:
-----
cluster1
-----
6
[2, 4, 10, 12, 3, 11]
cluster2
-----
3
[20, 30, 25]
  
```

Fig. 2: 2-clusters for Original Data

```

Clustered values are:
-----
cluster1
-----
6
[8, 12, 24, 28, 10, 26]
cluster2
-----
3
[44, 64, 54]
  
```

Fig. 3: 2-clusters for Transformed data

```

Clustered values are:
-----
cluster1
-----
6
[2, 4, 10, 12, 3, 11]
cluster2
-----
3
[20, 30, 25]
  
```

Fig. 4: 2-clusters for Normalized data

```

Clustered values are:
-----
cluster1
-----
3
[2, 4, 3]
cluster2
-----
3
[10, 12, 11]
cluster3
-----
3
[20, 30, 25]
  
```

Fig. 5: 3-clusters for Original Data

Table 5 and Figure 10 discuss the relation in the computation of our data set as original data, Transformed data and normalized data.

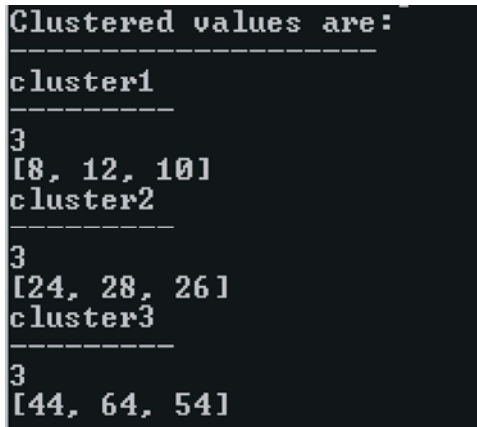


Fig. 6: 3-clusters for Transformed Data

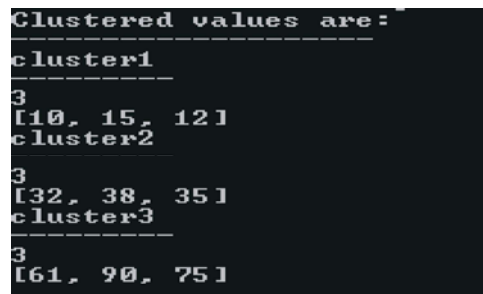


Fig. 7: 3-clusters for Normalized data

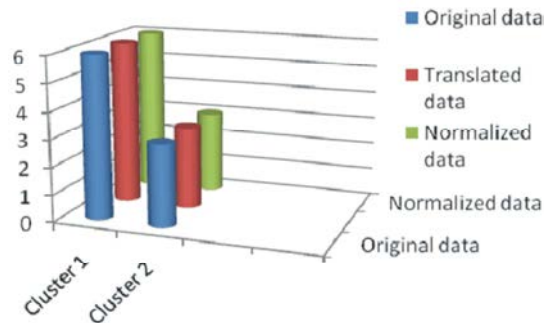


Fig. 8: Comparison Graph for 2-clusters

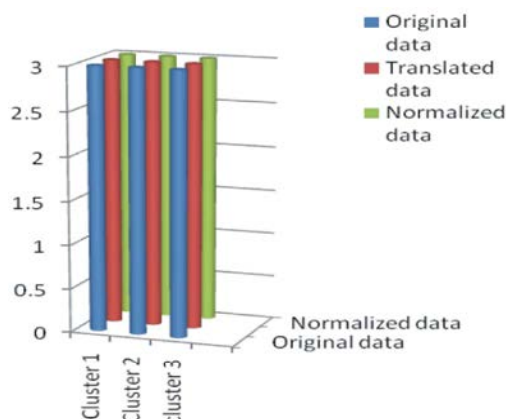


Fig. 9: Comparison Graph for 3-clusters

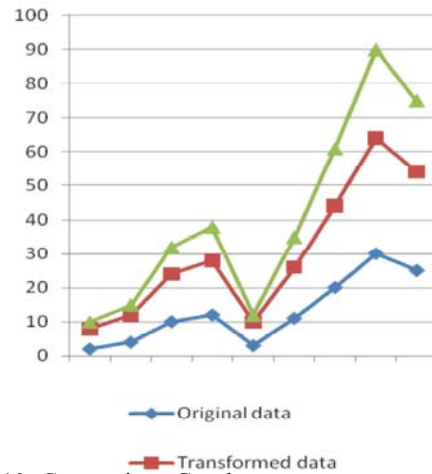


Fig. 10: Comparison Graph

From the above graph it is evident that the inter scalar distance between the elements is maintained the same in modified data as in the original data. It can also be concluded 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.

## CONCLUSION

In this work we have used a hybrid approach to preserve data privacy using a geometrical data transformation in the first step followed by normalization technique. From our experimental results, it is evident that the total number of elements in the clusters remains the same with the original and the modified data. This approach transforms the original data to privacy-preserved data maintaining the inter-relative distance among the data. Thus we have succeeded in achieving both accuracy and privacy. In future, this work can be extended with the use of other normalization techniques such as z-score normalization and decimal scaling in the second step.

## REFERENCE

1. Jiawei Han and Micheline Kamber, 2006. Data Mining-Concepts and Techniques, 2nd. Edition. San Francisco: Morgan Kaufmann Publishers.
2. 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.

3. 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.
4. 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*.
5. Likun Liu, Kexing Yang, Liang Hu and Liana Li, 2012. "Using Noise Addition Method Based on Pre-Minig to Protect Health care Privacy", *CEAI*, 14(2): 58-64.
6. Zhang Guo-rong, 2012. "An Effective Data Transformation Approach for Privacy Preserving Similarity Measurement", In *Proceedings 9th International Conference on Fuzzy Systems and knowledge Discovery (FSKD)*
7. Donald D. Hearn and M. Pauline Baker, 2011. *Computer Graphics*, 2nd. Edition. Pearson Publishers.
8. UCI Data Repository: <http://archive.ics.uci.edu/ml/datasets.html>.