Indian Journal of Science and Technology

Achieving Privacy in Data Mining Using Normalization

G. Manikandan[1*], N. Sairam[2], S. Sharmili[3] and S. Venkatakrishnan[4]

[1]Assistant Professor, School of Computing, SASTRA University, Thanjavur, India – 613401;
manikandan@it.sastra.edu
[2]Professor, School of Computing, SASTRA University, Thanjavur, India – 613401; sairam@cse.sastra.edu
[3, 4] Student, School of Computing, SASTRA University, Thanjavur, India – 613401;
sharmilisrinivasan@gmail.com,venkatakrishnan.sesh@gmail.com

Abstract

To extract the previously unknown patterns from a large data set is the ultimate goal of any data mining algorithm. Some private or confidential information may be revealed as part of data mining process. In this paper we use min-max normalization approach for preserving privacy during the mining process. We sanitize the original data using min-max normalization approach before publishing. For experimental purpose we have used k-means algorithm and from our results it is evident that our approach preserves both privacy and accuracy.

Keywords: Accuracy, Clustering, K-Means, Min-Max Normalization, Privacy.

1. Introduction

In today`s fast growing world, the amount of data being created and processed grows exponentially day by day. Organizations and companies are 'mining' huge data to draw conclusions that aid their decision making tasks. At times, there is also sharing of data among research institutions and companies for research analysis and to improve the quality of decisions respectively. The shared data may include confidential information such as medical records of individuals, criminal records, companies' financial records and so forth. Thus data sharing proves to be a threat to privacy of sensitive data of an individual or a company and their autonomy. Here comes the role of privacy preserving techniques which prevent the data from being leaked or misused, at the same time providing accurate mining results.

A number of privacy preserving algorithms have been proposed and are in use today. In this paper, we propose a new method using min-max normalization for preserving data during data mining. In general, min-max normalization is used as a pre-processing step in data mining for transformation of data to a desired range. Our objective is to use it for preserving privacy during data mining. We use K-means clustering to validate the proposed approach and verify for accuracy.

The rest of the paper is organized as follows: Section 2 provides an overview of literature works carried out in clustering techniques; Section 3 elaborates the implementation of min-max normalization and K-mean clustering techniques in our proposed system. Experimental results and simulations are tabulated and compared in Section 4 and finally in Section 5, we arrive to an overall conclusion from our work.

2. Literature Review

A set of hybrid geometrical data transformations namely HDTTR and HDSTR were used for clustering categorical data [1].

** Corresponding author:*
G. Manikandan (manikandan@it.sastra.edu)

Karthikeyan et al. [2] proposed a method for achieving privacy using fuzzy logic. For demonstrating their approach they have used s-shaped membership function for data sanitization. It is also proved that the clustering algorithm takes lesser number of iteration for clustering process with this modified data.

A shearing based data transformation approach was proposed by Manikandan et al. [3] for achieving privacy. Sheared data depends on the noise value. If the noise is more the user may easily identify that the data is a modified one and not original.

Liu et al. [4] proposed new Noise addition methods to protect health care privacy based on pre-mining.

A new protocol for clustering in a distributed environment based on additive sharing schemes instead of using homomorphic encryption with the objective to minimize computation and communication cost was proposed [5].

A novel algorithm for finding global frequent item sets with minimal communication overhead and time complexity was proposed by Rajalakshmi and Purusothaman [6].

3. Proposed System

In this paper we put forward an approach for privacy preserving using min-max Normalization.

3.1 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 [new_min_A,new_max_A], and the computation is given by

$$\frac{v - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new_{min_A}$$

where v' is the new value in the required range.

The advantage of Min-Max normalization is that it preserves the relationships among the original data values [7].

Table 1 is the sample data set used for experiment. Table 2 is the corresponding normalized values for the 'Age' attribute after applying min-max normalization.

The steps involved in our approach can be summarized in the form of a procedure as shown below. Figure 1 shows the flow diagram for the proposed system

Procedure

Step 1: Client will request the data from the Coordinator.

Step 2: Coordinator identifies the sensitive data in the data set.

Step 3: Sensitive Data are modified using Min-Max Normalization Process and the sanitized data is returned to the client.

Step 4: Client Uses K-means algorithm for clustering process.

Table 1. Original data

Sl.No	Name	Age	Gender
1	Anand	2	M
2	Abi	10	F
3	Mathi	20	F
4	Naveen	25	M
5	Sharmili	12	F
6	Venkat	30	M
7	Sai	20	M

Table 2. Normalized data for age attribute

Sl.No	Name	Age	Gender
1	Anand	10	M
2	Abi	33	F
3	Mathi	62	F
4	Naveen	76	M
5	Sharmili	39	F
6	Venkat	90	M
7	Sai	62	M

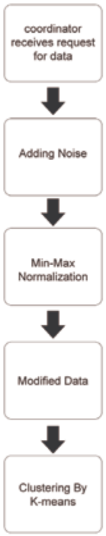


Figure 1. Flow Diagram for proposed system.

3.2 Clustering Technique

In our methodology, in order to check for the effectiveness of min-max normalization on the data partitioning techniques, we used K-means clustering algorithm. In K-means, the objects are clustered based on attributes into 'n' number of clusters where 'n' is a positive integer. The central idea of this Clustering is to minimize the sum of squares of the distance between data and corresponding cluster centroid in that data set. The clustering process must be carried out until it gets stabilized. Then, the objects are grouped based on the inter-relative distance between each object and the centroid. Figure 2 provides the flow chart of our clustering process.

4. Simulations and Results

In this paper, we have used min-max normalization to achieve privacy and accuracy during data mining and accuracy is tested using K-means clustering. Here the computations for min-max normalization of sample data, k-means clustering and effectiveness calculations are carried out in Java.

We have also tested the efficiency of our approach on a real-time dataset, "adult-dataset" from UCI data repository [8].

This data set comprises of 32561 records with 12 attributes namely age, work class, education, marital-status, occupation, relationship, race, sex, capital-gain and capital-loss, hours-per-week and native country. For experimental purpose, we used only age as the key attribute to carry out normalization in our work.

The clustering of data before and after normalization for 2-clusters is given in the figures Figure 3 and Figure 4 respectively. Similarly, those for 3-clusters are provided in the figures Figure 5 and Figure 6.

Table 3 and Table 4 describe the clustering of data before and after min-max normalization for 2-clustering and 3-clustering respectively.

Comparison of clustering prior and post normalization are shown in the figures Figure 7 and Figure 8 respectively.

The figures and tables clearly shows that clusters of data before and after applying min-max normalization remains the same.

Table 5 and Figure 9 summarises the comparisons among Fuzzy S-Shaped approach, Shearing Noise addition approach and our proposed min-max normalization approach.

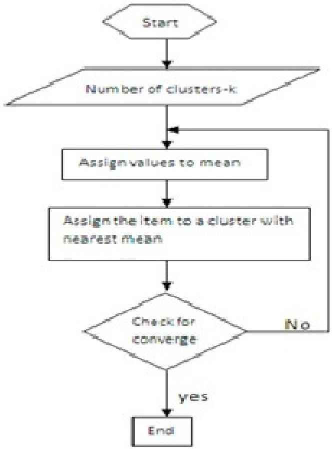
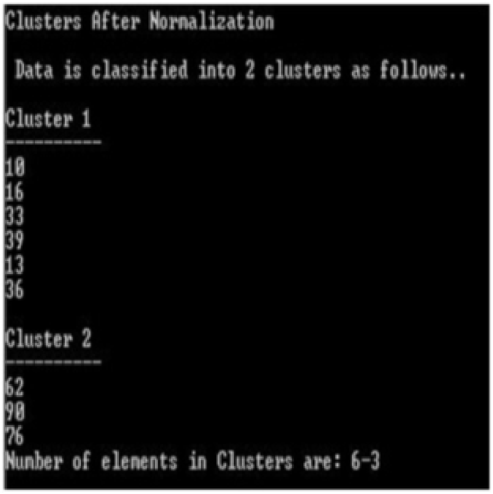
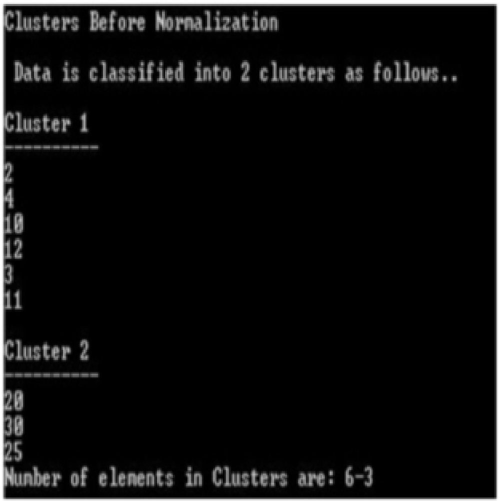


Figure 2. Clustering Process.



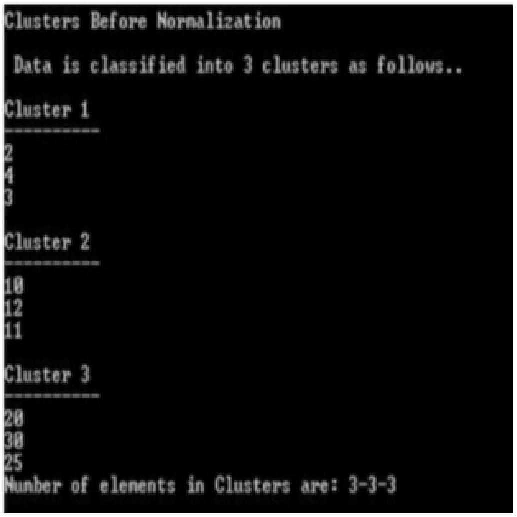
```
Clusters After Normalization

 Data is classified into 2 clusters as follows..

Cluster 1
----------
10
16
33
39
13
36

Cluster 2
----------
62
90
76
Number of elements in Clusters are: 6-3
```

Figure 3. Snapshot for 2-clusters- After Normalization.



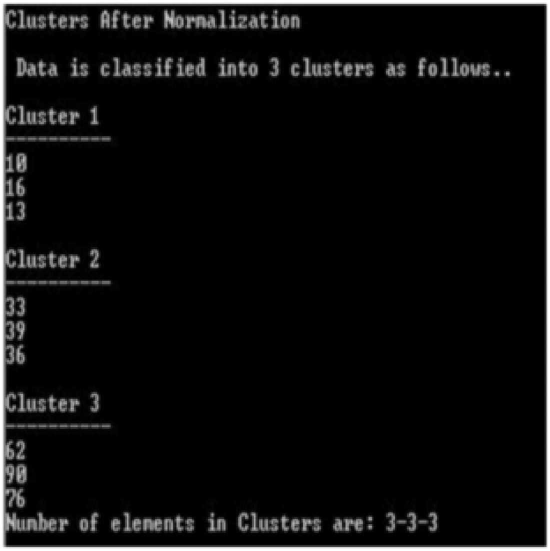
```
Clusters Before Normalization

 Data is classified into 2 clusters as follows..

Cluster 1
----------
2
4
10
12
3
11

Cluster 2
----------
20
30
25
Number of elements in Clusters are: 6-3
```

Figure 4. Snapshot for 2-clusters- Before Normalization.



```
Clusters Before Normalization

 Data is classified into 3 clusters as follows..

Cluster 1
----------
2
4
3

Cluster 2
----------
10
12
11

Cluster 3
----------
20
30
25
Number of elements in Clusters are: 3-3-3
```

Figure 5. Snapshot for 3-clusters- Before Normalization.



```
Clusters After Normalization

 Data is classified into 3 clusters as follows..

Cluster 1
----------
10
16
13

Cluster 2
----------
33
39
36

Cluster 3
----------
62
90
76
Number of elements in Clusters are: 3-3-3
```

Figure 6. Snapshot for 3-clusters- After Normalization.

Table 3. Results for 2-clusters

K = 2	Cluster 1	Cluster 2
Before Normalization	{2,4,10,12,3,11}	{20,30,25}
After Normalization	{10,16,33,39,13,36}	{62,90,76}

Table 4. Results for 3-clusters

K = 3	Cluster 1	Cluster 2	Cluster 3
Before Normalization	{2,4,3}	{10,12,11}	{20,30,25}
After Normalization	{10,16,13}	{33,39,36}	{62,90,76}

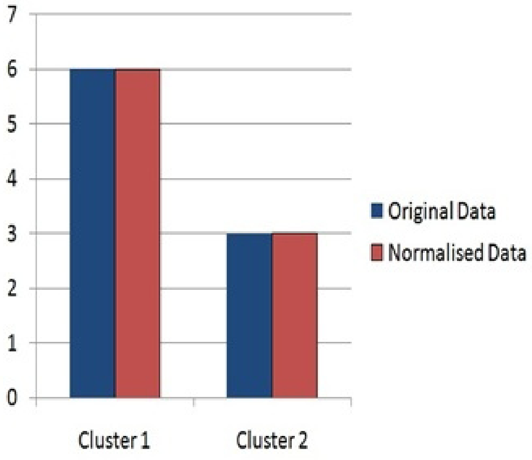


Figure 7. Comparison Graph for 2-clusters.

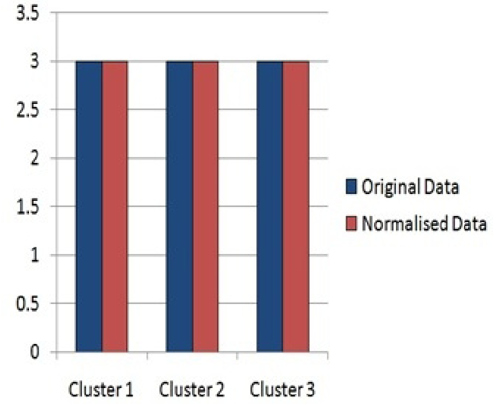


Figure 8. Comparison Graph for 3-clusters.

Table 5. Comparison tabulation

Original Data	Fuzzy S-Shaped	Shearing Noise (=10)	Min-Max Normalization
2	0	22	10
4	0.0102	44	16
10	0.1632	110	33
12	0.2551	132	39
3	0.0025	33	13
20	0.7449	220	62
30	1	330	90
11	0.2066	121	36
25	0.9362	275	76

From the above comparisons we observe that data transformation approach based on shearing, scales the values and scatters them over a large range. On the other hand,

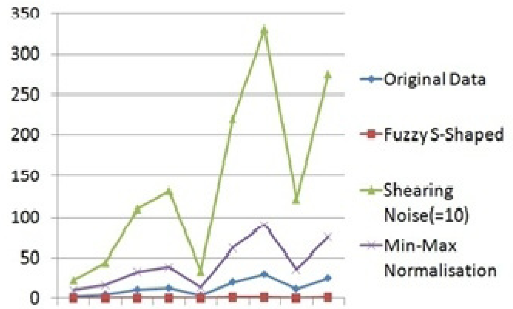


Figure 9. Comparison Graph.

Fuzzy approach based on S-Shaped membership function narrows down the range of values to [0, 1]. The produced results of both the approaches prove evidently the duplication of data for privacy preservation.

Our approach overcomes this limitation as the normalized values lie in the same range as the actual range of the attribute. Thus, the distortion of data for sake of privacy will not be revealed to the analyst or data-users, at the same time preserving privacy.

5. Conclusion

In this paper we have dealt with min-max normalization based data transformation to preserve data privacy. This approach transforms the original data to privacy-preserved data maintaining the inter-relative distance among the data. Experiments have proven that performing k-means clustering on the distorted data produces same clustering results as original data. Thus we have succeeded in achieving both accuracy and privacy. We have tested the technique for numerical data set. The future scope of this paper is to extend the same over categorical data.

6. References

1. Rajalaxmi R R, and Natarajan A M (2008). An effective data transformation approach for privacy preserving clustering, Journal of Computer Science, vol 4(4), 320–326.
2. Karthikeyan B, Manikandan G et al. (2011). A fuzzy based approach for privacy preserving clustering, Journal of Theoretical and applied information Technology, vol 32(2), 118–122.
3. Manikandan G, Sairam N et al. (2012). Privacy preserving clustering by shearing based data transformation, Proceedings of International Conference on Computing and Control Engineering.
4. Liu L, Yang K et al. (2012). Using noise addition method based on pre-mining to protect health care privacy, Journal of Control Engineering and Applied Informatics, vol 14(2), 58–64.
5. Doganay M, Pederson T et al. (2008). Distributed privacy preserving k-means clustering with additive secret sharing, PAIS '08 Proceedings of the International Workshop on Privacy and Anonymity in Information Society, 3–11.
6. Rajalakshmi M, and Purusothaman T (2011). Privacy preserving distributed data mining using randomized site selection, European Journal Of Scientific Research, vol 64(2), 610–624.
7. Han J, and Kamber M (2006). Data mining-concepts and techniques, 2nd Edn. San Francisco: Morgan Kaufmann Publishers.
8. Available from http://archive.ics.uci.edu/ml/datasets.html UCI Data Repository.