A Study on Normalization Techniques for Privacy Preserving Data Mining

C.Saranya[#1], G.Manikandan[*2]
[#]Computer Science & Engineering, School of Computing,
SASTRA UNIVERSITY, Tirumalaisamudram, Thanjavur – 613401. Tamilnadu, India.
[1]c.saranya213@gmail.com
[*]Computer Science & Engineering, School of Computing,
SASTRA UNIVERSITY, Tirumalaisamudram, Thanjavur – 613401. Tamilnadu, India.
[2] manikandan@it.sastra.edu

***Abstract* - Data mining is a prevailing technique which extracts the unfamiliar appealing patterns from large data sets. The extracted facts are utilized in various domains like marketing, weather forecasting, and medical diagnosis. It is very vital that the data gets exposed when the organizations start sharing the data for the mining process and privacy may be breached. Privacy is becoming a more and more significant issue in many data mining applications. Privacy preserving techniques gives a new track to solve this problem. It gives legitimate data mining outcomes without edifying the original data values and thus guarantees privacy as well as accuracy. In this paper we have analyzed the use normalization techniques in achieving privacy. We have compared the results of these techniques and from the experimental outcome it can be concluded that Min-Max normalization have minimum misclassification error.**

Keyword- Accuracy;Clustering;K-Means;Normalization; Privacy.

I.INTRODUCTION

Data Mining extracts the unknown information from a hefty heterogeneous data source which contains a large amount of private and sensitive data. To avoid the privacy leakage of data and to preserve the privacy as well as accuracy we use a special scheme called privacy preserving data mining (PPDM).

In the recent years PPDM plays an imperative role in data mining. It is the study of achieving some data mining goals without scarifying the privacy of the individuals. How to mine the patient's personal information? without violating his privacy in an ongoing research problem in this area. In the recent past number of privacy preserving methods has been proposed and analyzed in various dimensions.

In this paper we have analyzed the use of normalization techniques like Min-Max normalization, Z-Score normalization and Decimal Scaling methods with respect to privacy and accuracy. K-means Clustering algorithm is applied to the original and the tailored data to verify the effectiveness and the correctness of our proposed approach.

II. LITERATURE SURVEY

In [1] a set of hybrid data transformation is used to address the privacy problem. Misclassification error is used to measure the effectiveness of the clustering method in this work.

A new approach is proposed to preserve the sensitive information using fuzzy logic. Data can be sanitized by using a suitable membership function. Fuzzy concept is the extension of generic set theory. The limitation of this approach is that it maps all the values in the scale 0-1. From the sanitized data the user may easily infer that the data is not the original one. [2]

In [3] min-max normalization technique is used for preserving privacy during the mining process. The original data is sanitized using min-max normalization approach before publishing. Experiments were performed only with the numerical data.

Privacy can be realized by constructing a neural network for a simple linear transformation function. The efficiency of a neural network depends on the data set used for training the network. In [4] the network is trained with the original data itself to improve the efficiency.

In a distributed environment privacy can be achieved using various geometrical functions. The output of this transformation functions depends on the noise value. Normalization is done after geometrical transformation process to map the data to a uniform scale. [5]

III.PROPOSED WORK

The purpose of Normalization techniques is to map the data to a diverse scale. Various types of normalization techniques are available in the literature. In this paper we have compared three normalization techniques namely Min-Max, Z-Score and Decimal Scaling normalization.

Min-Max Normalization:

Min-max normalization performs a linear alteration on the original data. The values are normalized within the given range. For mapping a v value, of an attribute A from range [min_A,max_A] to a new range [new_min_A,new_max_A], the computation is given by,

$$\frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

where v' is the new value in the required range The benefit of Min-Max normalization is that all the values are annealed within certain range.

Z-Score Normalization:

Z score normalization, also called as Zero mean normalization. Here the data is normalized based on the mean and standard deviation. Then the formula is,

$$d' = \frac{d - mean(P)}{std(p)}$$

Where Mean(p) = sum of the all attribute values of P

Std(P)=Standard deviation of all values of P

*Decimal Scale Normalization***:**

Decimal Scale Normalization based on the movement of decimal point of value of attribute. The decimal point numbers are moved depends on the maximum absolute values of attribute. The decimal Scale normalization formula is,

$$d' = \frac{d}{10^m}$$

where, m is the smallest integer that max (| d' |)<*1*.

The original data is shown is shown in Table 1(a).The age attribute is the sensitive attribute and it is normalized. Table 1(b) shows the output of Min-Max normalization applied to age attribute. Table 1(c) and 1(d) depicts the normalized values based on Z-score and decimal scaling.

TABLE I(a) Original Data

S.No	Name	Age	Gender
1	Amu	4	F
2	Vijay	12	M
3	Divya	20	F
4	Udhaya	26	F
5	Ram	30	M

TABLE I(b) MinMax NormalizedData

S.No	Name	Age	Gender
1	Amu	10	F
2	Vijay	34	M
3	Divya	59	F
4	Udhaya	77	F
5	Ram	90	M

TABLE I(c) Z -Score Normalized Data

S.No	Name	Age	Gender
1	Amu	1.32	F
2	Vijay	0.56	M
3	Divya	0.18	F
4	Udhaya	0.75	F
5	Ram	1.13	M

TABLE I(d) Decimal Scale Normalized Data

S.No	Name	Age	Gender
1	Amu	0.04	F
2	Vijay	0.12	M
3	Divya	0.2	F
4	Udhaya	0.26	F
5	Ram	0.3	M

K-Means Clustering:

Clustering is an unsupervised learning technique which groups the similar objects into appropriate clusters. Flow chart in Figure 1 summarizes the steps involved in K-means clustering algorithm.

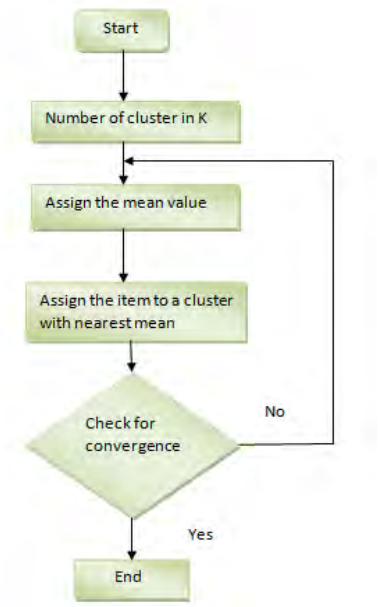

Fig 1. K -Means Clustering Algorithm

IV. SIMULATIONS AND RESULTS

In this paper we have implemented various normalization techniques to achieve privacy in data mining. We have also computed the effectiveness and of these techniques using K means Clustering algorithm. The following snapshots are based on the outcome of K-mean clustering algorithm for 2 clusters with the original and Normalized data.

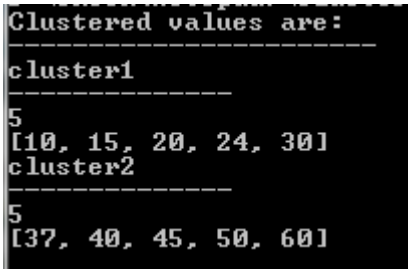

Figure.3 2-cluster for original data

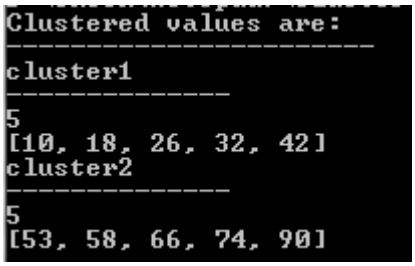

Figure.4 2-cluster for Min-Max data

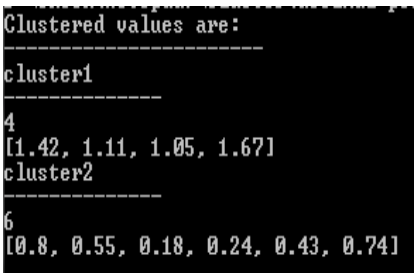

Figure.5 2-cluster for Z-Score data

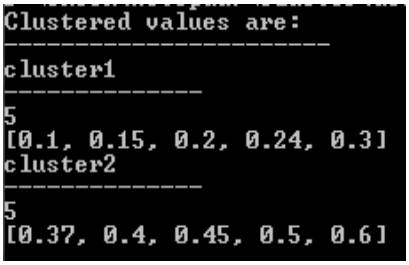

Figure.6 2-cluster for Decimal Scale data

Table 2 describes the clustering result of original and normalized data. Table 3 and Fig.2 summarises the comparisons among Original and the Normalized Data.

Table II -cluster Result

K is 2	1- Cluster	2- Cluster
Original Data	[10,15,20,24,30]	[37,40,45,50,60]
MinMax Data	[10,18,26,32,42]	[53,58,66,74,90]
Z-Score Data	[1.42,1.11,1.05,1.67]	[0.8,0.55,0.18,0.24,0.43,0.74]
Decimal Scale Data	[0.1,0.15,0.2,0.24,0.3]	[0.37,0.4,0.45,0.5,0.6]

Table III- Comparison Table

	Normalization process		
Original Data	Min-Max	Z-Score	Decimal Scale
10	10	1.42	0.1
15	18	1.11	0.15
20	26	0.80	0.2
24	32	0.55	0.24
a30	42	0.18	0.3
37	53	0.24	0.37
40	58	0.43	0.4
45	66	0.74	0.45
50	74	1.05	0.5
60	90	1.67	0.6

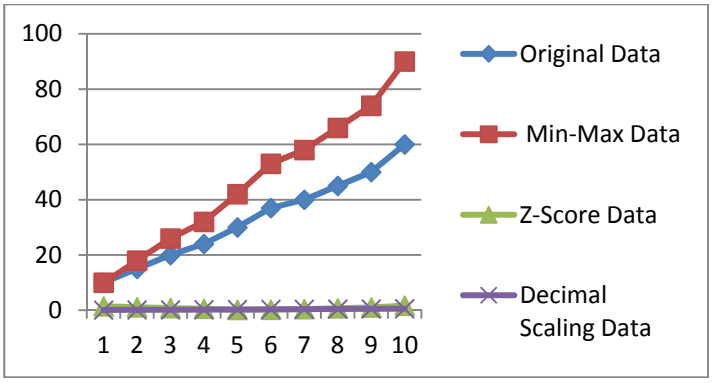

Figure.2. Comparison Graph

V. CONCLUSION

In this work, we have analysed the use of normalization techniques to preserve data privacy. We carried out all the three normalization techniques on a sample data set containing 10 elements with values 10, 15, 20, 24, 30, 37, 40, 45, 50, 60 are used and from the experiments and results, it is clearly evident that Min-Max normalization has less number of misclassification errors when compared to Z-Score and Decimal Scaling.

REFERENCES

[1] G.Manikandan,N.Sairam, R.Sudan and B.Vaishnavi "Shearing based Data Transformation Approach for Privacy Preserving Clustering " Third International Conference on Computing,Communication and Networking Technologies (ICCCNT-2012), SNS College of Engineering,Coimbatore,July 26-28,2012

[2] B.Karthikeyan,G.Manikandan,Dr.V.Vaithiyanathan,A Fuzzy Based Approach for Privacy Preserving Clustering in Journal of Theoretical and applied information Technology , Vol 32(2), October 2011,118-122(Scopus)

[3] G.Manikandan,N.Sairam,S.Sharmili,S.Venkatakrishnan , Data Masking – A few new Techniques – international conference on research and development prospects on engineering and technology(ICRDPET -2013), E.G.S Pillay Engineering college, Nagapattinam , march 29-30,2103.

[4] G.Manikandan,N.Sairam,S.Sharmili,S.Venkatakrishnan , Achieving Privacy in Data Mining Using Normalization in Indian Journal of Science and Technology, Vol 6(4) , 18-April 2013,4268-4272.

[5] G.Manikandan,N.Sairam,S.Jayashree,C.Saranya , Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation in Middle East Journal Of Scientific Research , Vol 14(1) , 20-April 2013,107-111.