Middle-East Journal of Scientific Research 14 (1): 107-111, 2013
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2013.14.1.9313

Achieving Data Privacy in a Distributed Environment Using Geometrical Transformation

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

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

Abstract: Advances in data acquirement methods have resulted in gathering and storing cosmic quantities of data. For communal profit, these data are shared among various organizations for investigative purposes. In various astonishing circumstances this data sharing may reveal some concealed information thus raising privacy concerns. This paper introduces a proficient privacy-preserving technique for a distributed environment. Prior to data sharing at each site a geometrical transformation is used to modify the original data which is followed by a normalization process. To verify the untried results k-means clustering algorithm is used and it is apparent from our experimental outcome that our approach preserves privacy and also ensures accuracy in a distributed environment.

Key words Translation · Scaling · Shearing · Clustering · Normalization

INTRODUCTION

Because of the swift progress in the technologies like networking; hardware and software, there is remarkable growth in the amount of data that can be collected by an organization. Data mining is a dominant tool 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. Privacy preserving data mining gives legitimate data mining results without learning the original data values and thus guarantees privacy for the sensitive data housed in a data warehouse. The payback of data mining can be enjoyed, without compromising the individual's privacy. The original data is customized or a procedure is used in such a way that concealed data and confidential knowledge remain private even after the mining process is carried out.

In this manuscript we suggest a new method that ensures the data privacy in a distributed environment. When there is a need to share the data, the data providers transform the data by applying a geometrical transformation which is basically a noise addition framework. We make use of k-means clustering technique to corroborate the results to substantiate the accuracy of the proposed work.

The organization of this paper is as follows: Section II provides an overview of literature works carried out in privacy techniques; Section III elaborates the implementation of various geometrical transformations used in our work. Experimental results, Simulation snapshots are summarized in Section IV and finally Section V provides an overall conclusion of our work.

Literature Review: [1] use additive secret sharing scheme to cluster the vertically partitioned data in a distributed environment with a less communication overhead.

For categorical data clustering [2] has proposed a data transformation approach which uses a set of geometrical data transformation functions like HDTTR and HDSTR.

Usually shearing is used to change the shape of an object in geometrical transformation. [3] has proved that the data privacy can be achieved using shearing data transformation. The limitation of this approach is the transformed data depends on the noise values used for shearing.

Fuzzy concept can be employed to achieve privacy. The data is modified using a appropriate membership function. [4] has utilized an s-shaped membership function for data sanitization.

Corresponding Author: G. Manikandan, School Of Computing, SASTRA University, Thanjavur, India, 613402

Transformations like translation, scaling can be clubbed with shearing and it can be used to transform the data by changing their sequence. This hybrid approach for data perturbation is successfully demonstrated by [5].

Use of Elliptic Curve Cryptography (ECC) along with randomized site selection to find the global frequent itemsets in a distributed environment with minimal communication cost was proposed by [6].

[7] have proposed a framework that integrates randomized data perturbation technique with cryptography to transform the original data in a distributed environment.

Proposed System: In this work, we suggest a new approach using geometrical data transformation for preserving privacy in a distributed environment. In a distributed scenario the data is scattered among multiple locations. In each site a transformation like translation, scaling or shearing can be used to modify the data. As a result the data is scaled to different ranges depending on the operation and the noise. Subsequent to the transformation operation a normalization process is employed in different sites to map the data to a uniform scale. Here we model age attribute as two-dimensional object (where the value of x and y are assumed to be the same i.e x=y) so that it can be resized or repositioned by applying basic transformations.

Shearing Based Transformation: Shearing Distorts the Shape of an Object So That the New Shape Appears as If the Object Were Composed of Internal Layers

$$X'=X+ (Sh_x*X)$$

where X is the original data , Sh_x is the noise, X' is the Sheared data.

Translation Based Transformation: When translation is applied to an object it is repositioned from one coordinate location to the other in a straight line path. This is done by adding translation distance to the original coordinate. To translate a point x to a new location x' the following equation is used

$$X'=X+T_x$$

Where X is the original data, Tx is the noise to translate the original data, X' is the translated data.

Table 1(a) Original data

S.NO	NAME	AGE	GENDER
1	SRI	3	F
2	RAJA	11	M
3	SARANYA	19	F
4	CHANDRU	27	M
5	GAYATHIRI	31	F

Table 1(b) Sheared data

S.NO	NAME	AGE	GENDER
1	SRI	9	F
2	RAJA	33	M
3	SARANYA	57	F
4	CHANDRU	81	M
5	GAYATHIRI	93	F

Table 2(a) Original data

S.NO	NAME	AGE	GENDER
1	SRI	3	F
2	RAJA	11	M
3	SARANYA	19	F
4	CHANDRU	27	M
5	GAYATHIRI	31	F

Table 2(b) Translated data

S.NO	NAME	AGE	GENDER
1	SRI	5	F
2	RAJA	13	M
3	SARANYA	21	F
4	CHANDRU	29	M
5	GAYATHIRI	33	F

Table 3(a) Original data

S.NO	NAME	AGE	GENDER
1	SRI	3	F
2	RAJA	11	M
3	SARANYA	19	F
4	CHANDRU	27	M
5	GAYATHIRI	31	F

Table 3(b) Scaled data

S.NO	NAME	AGE	GENDER
1	SRI	6	F
2	RAJA	22	M
3	SARANYA	38	F
4	CHANDRU	54	M
5	GAYATHIRI	62	F

Scaling Based Transformation: To change the size of an object scaling can be used. This can be achieved by multiplying the coordinate values by scaling factors. The transformed coordinate X' can be obtained by multiplying the original coordinate X with a scaling factor Sx.

Table 4: Original and Normalized data at different sites

	SITE-1 Noise=2		SITE-2 Noise=2		SITE-3 Noise=2	
Original Data	Shearing	Normalized Data	Translation	Normalized Data	Scaling	Normalized Data
6	18	10	8	10	12	10
9	27	16	11	16	18	16
14	42	27	16	27	28	27
16	48	32	18	32	32	32
17	51	34	19	34	34	34
19	57	38	21	38	38	38
23	69	47	25	47	46	47
26	78	54	28	54	52	54
35	105	74	37	74	70	74
55	126	90	44	90	84	90

$$X'=X*S_x$$

Where X is the original data, Sx is the noise to translate the original data, X' is the scaled data.

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 $[new\_min_A,new\_max_A]$, 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 advantage of Min-Max normalization is that it preserves the relationships among the original data values [8].

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 submits a request for data to the Coordinator.
Step 2: Coordinator identifies the potential data owners.
Step 3: Coordinator generates a random noise and the data range.
Step 4: Coordinator sends a request for data along with the noise and data range to the identified Data owners.
Step 5: At each site the data is modified using a geometrical transformation and the modified data is normalized.

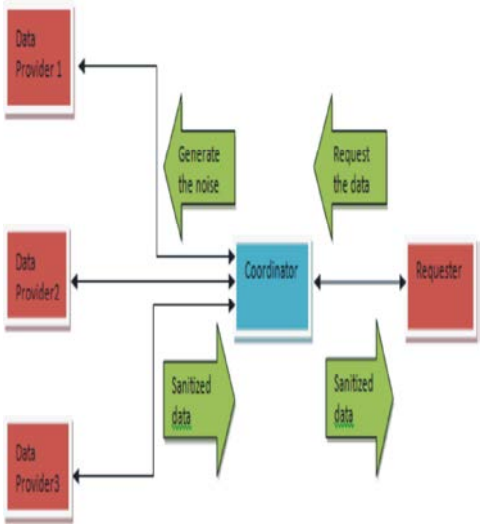

Fig. 1: Flow Diagram for proposed system

Step 6: Data owner forwards the normalized data to the coordinator.
Step 7: Coordinator transmits the data to the user.

Simulation and Results: In this paper, we have used geometrical transformation in each site followed by a normalization process to achieve privacy and accuracy during data mining in a distributed environment. The accuracy is tested using K-means clustering. Here the computations for the synthetic data set (6,9,14,16,17,19,23,26,35,42), k-means clustering and effectiveness calculations are carried out in Java. For experimental purpose we presume that there are three sites in a distributed environment. In the first site we apply shearing, in the second site translation and scaling in the third site with the noise value as 2 .The original data, transformed data at different sites and the equivalent normalized data at each site is shown in the Table 4.

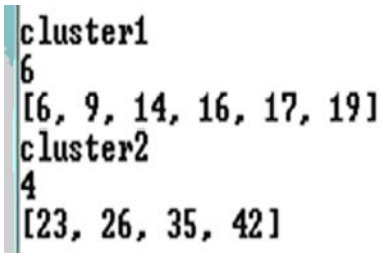

Fig. 2: original data with 2 Cluster

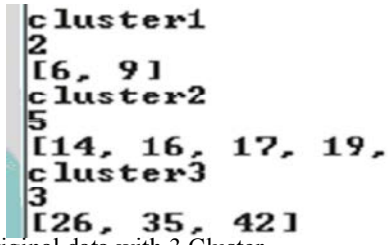

Fig. 3: original data with 3 Cluster

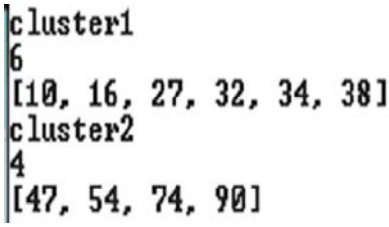

Fig. 4: original data with 2 Cluster

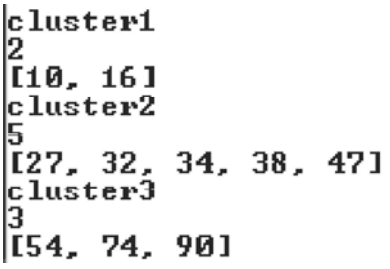

Fig. 5: original data with 3 Cluster

From the Table 4 it is evident that irrespective of the transformation function used the ensuing normalized data is the same at all the sites. For verifying accuracy we have applied clustering to the original and the normalized data. The results conclude that the total number of elements remains the same in the resultant clusters. Figure (2) and (3) are the snapshots of the resultant clusters for the original data and the snapshots of the resultant clusters of the transformed data are shown in figure (4) and (5). Table 5 summarizes the contents of various clusters generated by K-means clustering algorithm.

Table 5: Clustering output

K=3	Cluster1	Cluster2	Cluster3
Original data	6,9	14,16,17,19,23	26,35,42
Normalized data	10,16	27,32,34,38,47	54,74,90

CONCLUSION

Data mining extract valuable patterns from hefty quantities of data stored in the data repository. The outcome of data mining process results in precious patterns which are used to support various decisions in different domains. But, such repositories also contain confidential and insightful information and the release of this personal information can cause momentous harm to data owner. In this paper we presented an approach in which data perturbation technique and normalization are integrated to provide better data quality and ensure individual privacy in a distributed environment. In Perturbation a small amount of noise is added to susceptible data such that the properties and the connotation of the original data are not changed. We obtain enhanced privacy because the result obtained is in perturbed form, so the privacy of original data is preserved by giving valid data mining results.

REFERENCES

1. Doganay, M., T. Pederson, Y. Saygin, E. Savas and A. Levi, 2008. Distributed Privacy Preserving Clustering with Additive Secret Sharing, In proceedings of the international workshop on privacy and anonymity in Information Society table, pp: 3-11.
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. Manikandan, G., and R. Sudhan Vaishnavi, 2012. Privacy Preserving Clustering By Shearing based Data Transformation, In Proceedings of International Conference on Computing and Control Engineering. ICCCE.
4. Karthikeyan, B., G. Manikandan and Dr. V. Vaithiyanathan, 2011. A Fuzzy Based Approach for Privacy Preserving Clustering, Journal of Theoretical and applied information Technology, 32(2): 118-122.
5. Manikandan, G., N. Sairam and R. Sudhan 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.
6. Rajalakshmi, M. and T. Purusothaman, 2011. Privacy Preserving Distributed Data Mining using Randomized Site Selection, European Journal Of Scientific Research, 64(2): 610-624.

7. Kamakshi, P. and Dr. A.Vinaya Babu, 2010. Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed Data, Journal of Computing, 2(4): 115-119.
8. Jiawei Han and Micheline Kamber, 2006. Data Mining-Concepts and Techniques, 2nd. Edition. San Francisco: Morgan Kaufmann Publishers.