A Novel Approach for Horizontal Privacy Preserving Data Mining

Hanumantha Rao Jalla and P.N. Girija

Abstract Many business applications use data mining techniques. Small organizations collaborate with each other to develop few applications to run their business smoothly in competitive world. While developing an application the organization wants to share data among themselves. So, it leads to the privacy issues of the individual customers, like personal information. This paper proposes a method which combines Walsh Hadamard Transformation (WHT) and existing data perturbation techniques to ensure privacy preservation for business applications. The proposed technique transforms original data into a new domain that achieves privacy related issues of individual customers of an organization. Experiments were conducted on two real data sets. From the observations it is concluded that the proposed technique gives acceptable accuracy with K-Nearest Neighbour (K-NN) classifier. Finally, the calculation of data distortion measures were done.

Keywords Horizontal privacy preserving · Walsh hadamard transformation · Data perturbation · Classification

1 Introduction

Explosive growth in data storing and data processing technologies has led to creation of huge databases that contains fruitful information. Data mining techniques are used to retrieve hidden patterns from the large databases. In distributed data

H.R. Jalla (✉)
Department of Information Technology, CBIT, Hyderabad, T.S, India
e-mail: hanucs2000@gmail.com

P.N. Girija
School of Computer and Information Sciences, UoH, Hyderabad, T.S, India
e-mail: pn_girija@yahoo.com

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 434,
DOI 10.1007/978-81-322-2752-6_9

mining a group of clients share their data with trusted third party. For example a health organization collects data about diseases of a particular geographical area that is nearby to the organization. To improve the quality of information and collaborate with other organizations, which benefits the participated clients to conduct their business smoothly. The Third party performs a knowledge based technique on the collected data from group of clients. In this scenario, clients can be grouped under three categories like honest, semi-honest and malicious. First category honest client always obey the protocol and will not alter the data or information. Second category, semi-honest client follows the protocol but tries to acquire the knowledge about other clients while executing. Third category, malicious client or unauthorized client always tries to access other's data and alter the information. To provide security of the data, this paper proposes a novel approach.

Many researchers address this problem by using cryptography, perturbation and reconstruction based techniques. This paper proposes an approach for Horizontal Privacy Preserving Data Mining (HPPDM), with combination of different transformations. Consider a Trusted Party (TP) among group of clients. Then TP communicate with other clients by using symmetric cryptography algorithm and assigns transformation techniques to each client. In this work, transformation techniques such as WHT, Simple Additive Noise (SAN), Multiplicative Noise (MN) and FIrst and Second order sum and Inner product Preservation (FISIP) are used. Proposed model is evaluated with data distortion and privacy measures such as Value Difference (VD) and Position Difference parameters like CP, RP, CK and RK.

This paper is organized as follows: Sect. 2 discusses about the Related Work. Section 3 focus on Transformation Techniques. Section 4 explains about Proposed Model. Section 5 discusses Experimental Results and in Sect. 6 Conclusion and Future Work discussed.

2 Related Work

Recently there has been lot of research addressing the issue of Privacy Preserving Data Mining (PPDM). PPDM techniques are mostly divided into two categories such as data perturbation and cryptographic techniques. In data perturbation methods, data owners can alter the data before sharing with data miner. Cryptography techniques are used in distributed data mining scenario to provide privacy to individual customers.

Let X be a numerical attribute and Y be the perturbed value of X. Traub [1] proposed SAN as $Y = X + e$ and MN as $Y = X * e$. Other perturbation methods such as Micro Aggregation (MA), Univariate Micro Aggregation (UMA) and

Multivariate Micro Aggregation (MMA) are proposed in [2–4]. Algorithm based Transformations are discussed in Sect. 3.

Distributed Data Mining is divided into two categories Horizontal and vertical in PPDM. Yao [5] introduced two-way Communication protocol using cryptography. Murat and Clifton proposed a secure K-NN classifier [6]. In [7] Yang et al. proposed a simple cryptographic approach i.e. many customers participate without loss of privacy and accuracy of a classifier. A frame work was proposed [8] which include general model as well as multi-round algorithms for HPPDM by using a privacy preserving K-NN classifier. Kantarcioglu and Vaidya proposed privacy preserving Naive Bayes classifier for Horizontally Partition data in [9]. Xu and Yi [10] discussed about classification of privacy preserving Distributed Data Mining protocols.

3 Transformation Techniques

3.1 Walsh Hadamard Transformation

Definition The Hadmard transform H_n is a $2^n \times 2^n$ matrix, the Hadamard matrix (scaled by normalization factor), that transforms 2^n real numbers X_n into 2^n real numbers X_k. The Walsh-Hadamard transform of a signal X of size $N = 2^n$, is the matrix vector product $X*H_n$. Where

$$H_N = \overset{n}{_{i=1}} \otimes H_2 = \underbrace{H_2 \otimes H_2 \otimes \ldots \otimes H_2}_{n}$$

The matrix $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\otimes$ denotes the tensor or kronecker product. The tensor product of two matrices is obtained by replacing each entry of first matrix by multiplying the first matrix with corresponding elements of second matrix. For example

$$H_4 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

The Walsh-Hadamard transformation generates an orthogonal matrix, it preserves Euclidean distance between the data points. This can be used in Image Processing and Signal Processing.

3.2 First and Second Order Sum and Inner Product Preservation (FISIP)

FISIP is a distance and correlation preservation Transformation [11].

Definition The matrix representation of a linear transformation can be written as $A = [A_i] = [A_1\ A_2 \ldots A_k]$, Additionally, A_i can be written as $A_i = [A_{im}]$. Then the transformation is called a FISIP transformation if A has following properties.

a. $\sum_{m=1}^{k} A_{im} = 1.$

b. $\sum_{m=1}^{k} A_{im}^2 = 1.$

c. $\sum_{m=1}^{k} A_{im}A_{jm} = 0,\ for\ i \neq j$

$$A^{[k]} = \left[a_{ij}\right], 1 \leq j \leq k, a_{ii} = \frac{2-k}{k}, a_{ij} = \frac{2}{k}$$

$$A^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, A^3 = \begin{bmatrix} \frac{-1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{-1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{-1}{3} \end{bmatrix}$$

3.3 Discrete Cosine Transformation (DCT)

DCT works on real numbers and gives following real coefficients:

$$f_i = \left(\frac{2}{n}\right)^{\frac{1}{2}} \sum_{k=0}^{n-1} \Lambda_k x_k \cos[(2k+1)i\Pi/2n]$$

where for k = 0 and 1 otherwise. These transforms are unitary and Euclidean distance between two sequences is preserved [12].

3.4 Randomization

Randomization is one of the simple approaches to PPDM. Randomization involves perturbing numerical data. Let X is a confidential attribute Y be a perturbed data

[1]. SAN is defined as$Y = X + e$, where e is random value drawn from a distribution with mean value zero and variance 1.

MN is defined as $Y = X * e$, e is a random value.

4 Proposed Model

This paper proposes a new approach for HPPDM as shown in Fig. 1. In this approach a group of clients select a Trusted Party who has capability to retrieve information from large data. Symmetric cryptography algorithm is used for communication between clients. Suppose client1 wants to collaborate with other client i.e., client2. Client1 sends a request to client2 for approval of collaboration. If client2 sends acceptance response both the clients choose their own transformation/modification techniques to modify the data.

This work focuses on perturbs numeric data. Numerical attributes are considered as a confidential attribute. Different transformations techniques are used to modify original data, transformations techniques discussed in Sect. 3. Both clients modify their original data using transformation techniques then modified data will be sent to the Trusted Party. TP decrypts modified data which is collected from clients and performs knowledge based technique. K-Nearest Neighbor (K-NN) as knowledge based technique.

Theorem *Suppose that T:* $R^n \rightarrow R^n$ *is a linear transformation with matrix A, then the linear transformation T preserves scalar products and therefore distance between points/vectors if and only if the associated matrix A is orthogonal.*

Here $T(X) = T_1(X_1) + T_2(X_2)$, T_1 and T_2 are linear transformations subsequently; T is also a linear Transformation. Whereas $X_1$and X_2 are the original data of the respective clients.

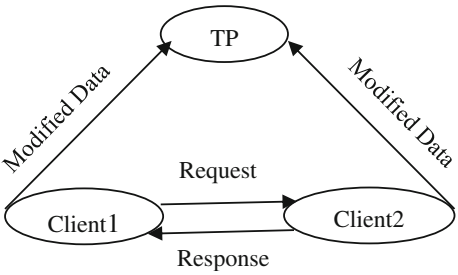

Fig. 1 Model for HPPDM

Proposed Algorithm

Client1:

```
Input: D1[M1][N]
Output: D1′[M1][N]

Begin
      If (N<2^n)  (n=0, 1, 2, 3…..)
            Append Columns with Zeros its nearest 2^n
      Endif
      A=WHT (N) // generate WHT matrix
      MD1= D1 * A;
End
```

Client2:

```
Input: D2 [M2][N]
Output: D′2[M2][N]

Begin
      Choose any one of the transformation technique discussed in section
3.
End
```

Trusted Party:

```
Input: D1 [M1][N] ,D2[M2][N]
Output: D [M][N]

Begin
      Step1: Collect modified data from client1
      Step2: Collect modified data from client2
      Step3: Merge both modified data sets which are received from client1
      and client2
      Step4: Perform knowledge based technique.
      Step5: Share knowledge i.e., obtained from step4 to both the clients.

End
```

Any data like parameters, keys and modified data that needs to be securely shared between the clients and TP must be ensure using Symmetric Cryptography Algorithms.

5 Experimental Work

Experimental work conducted on two real datasets Iris and WDBC collected from [13]. Assume that datasets taken as matrix format, row indicates an object and column indicates an attribute. Divide entire dataset into two parts numerical and categorical attributes. Among that only numerical attributes are considered and it is shared between the clients.

The distribution of the data is done in two methods. Generate a random value using random function which is considered as the percentage of the total records. In method 1: For client1 the data records are sent from record 1 to the percentage of

random value. And the remaining records are sent to client2. In method 2: first n records are leaving (n value change as per the dataset size) and then consider the random value to select next number of records to send for client 1. For client2 merge few records from first n records and remaining from the left out records of the whole data set and send them. If any client choose WHT as transformation technique data pre-processing is required if Number of Attributes are less than 2^n, n = 0, 1, 2, 3,… add number of columns to its nearest 2^n (Table 1).

K-NN is used as a classification technique from WEKA Tool [14]. While conducting experiments K value set to 3, 5, and 7. Tenfold cross validation is used when running the K-NN algorithm. In this paper, consider four combinations of linear transformations such as WHT-WHT, WHT-DCT, WHT-FISIP and WHT-SAN. Follow two methods in data distribution, select 4(2 values are below 50 and 2 values are above 50 to 100) random values in each method per a dataset. 35 and 135 records are skipped in IRIS and WDBC datasets respectively as per method 2 discussed above. Classifier results of IRIS data set shown in Tables 2, 3, 4 and 5. Classifier Accuracy of WDBC shown in Tables 6, 7, 8 and 9. IRIS original data gives 96.00 % using K-NN. Modified IRIS data gives acceptable accuracy on k = 7 using all methods. WDBC Original data set gives 97.18 and modified WDBC gives acceptable accuracy in all cases. Calculated different privacy measures VD, RP, CP, RK and CK from [15].

Calculated average values of all data distributions shown in Tables 10 and 11. The higher values of RP and CP and the lower value of RK and CK, and the analysis show more privacy is preserved [15].

Table 1 Data set description

Data set	No. of records	No. of attributes	No. of classes
IRIS	150	4	3
WDBC	569	31	2

Table 2 Accuracy on Iris data set using combination of WHT-WHT

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
28-72	96.00	96.00	96.00	96.67	96.00	94.67
45-65	96.00	96.00	96.00	96.67	96.00	94.67
65-35	96.00	96.00	96.00	96.67	96.00	94.67
71-29	96.00	96.00	96.00	96.67	96.00	94.67
51-49*	95.33	95.33	96	94	96.67	95.33
55-45*	95.33	96	955.33	96	96.67	96
70-30*	95.33	95.33	96	95.33	94.67	95.33
76-24*	95.33	94.67	95.33	95.33	96	95.33

*Skip first 35 records for data distribution

Table 3 Accuracy on Iris data set using combination of WHT-DCT

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
27-73	96.00	96.67	96.00	97.33	96.00	97.33
46-54	96.00	96.67	96.00	96.67	96.00	98
61-39	96.00	100	96.00	100	96.00	100
75-25	96.00	97.33	96.00	97.33	96.00	98
26-74*	96.00	98.67	96.00	98	96.67	98
45-55*	96.00	96.67	96.00	96.67	96.67	96.67
54-46*	94.00	98	95.33	98	95.33	98
58-42*	95.33	98	95.33	98	97.33	98

*Skip first 35 records for data distribution

Table 4 Accuracy on Iris data set using combination of WHT-FISIP

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
29-71	96.00	96.00	96.00	96.00	96.00	96.00
39-61	96.00	96.00	96.00	94.67	96.00	94.67
55-45	96.00	97.33	96.00	97.33	96.00	97.33
76-24	96.00	98.67	96.00	98	96.00	98
34-66*	95.33	97.33	94.67	97.33	96	97.33
47-53*	95.33	99.33	96.00	99.33	96.00	99.33
53-47*	94.00	98.67	96.00	98	96.00	98
57-43*	95.33	98	95.33	98.00	96.00	98

*Skip first 35 records for data distribution

Table 5 Accuracy on Iris data set using combination of WHT-SAN

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
25-75	96.00	96.00	96.00	95.33	96.00	94.67
34-66	96.00	94.00	96.00	94.00	96.00	94.67
58-42	96.00	99.33	96.00	99.33	96.00	99.33
79-31*	96.00	98.66	96.00	98.00	96.00	98.00
24-76*	95.33	96.00	96.67	96.00	96.67	96.00
50-50*	95.33	98.67	96.00	98.00	96.00	98.67
55-45*	95.33	98.00	96.00	98.00	97.33	97.33

*Skip first 35 records for data distribution

Table 6 Accuracy on WDBC data set using combination of WHT-WHT

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
17-83	96.83	92.44	97.01	92.79	97.18	92.44
35-65	96.83	92.44	97.01	92.79	97.18	92.61
50-50	96.83	92.44	97.01	92.79	97.18	92.61
68-32	96.83	92.44	97.01	92.79	97.18	92.61
35-65**	97.12	92.26	97.36	93.84	97.36	93.32
48-52**	97.18	92.61	96.83	92.97	97.01	93.49
59-41**	97.01	92.09	96.48	92.07	97.18	93.14

**Skip 135 records for data distribution

Table 7 Accuracy on WDBC data set using combination of WHT-DCT

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
31-69	96.83	57.82	97.01	61.68	97.18	62.21
49-51	96.83	93.67	97.01	94.02	97.18	94.20
63-37	96.83	93.49	97.01	93.32	97.18	92.44
79-21	96.83	92.79	97.01	92.97	97.18	92.79
19-81**	97.01	92.61	96.66	92.44	96.66	92.09
37-63**	96.66	92.97	97.01	92.79	97.01	92.26
69-31**	97.01	92.26	96.66	93.32	97.18	92.97
75-25**	97.01	92.26	97.01	93.32	96.83	92.79

**Skip 135 records for data distribution

Table 8 Accuracy on WDBC data set using combination of WHT-FISIP

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
24-76	96.83	93.32	97.01	93.67	97.18	93.84
36-64	96.83	93.49	97.01	93.84	97.18	93.84
65-35	96.83	93.32	97.01	92.97	97.18	93.32
83-27	96.83	93.49	97.01	93.32	97.18	92.79
30-70**	96.83	92.44	96.66	93.32	97.18	91.56
46-54**	96.83	92.44	97.18	91.91	97.18	91.56
65-35**	97.18	92.97	96.66	92.44	96.83	92.44
74-26**	96.84	93.67	96.83	93.84	96.83	93.67

**Skip 135 records for data distribution

Table 9 Accuracy on WDBC data set using combination of WHT-SAN

%split (client1-client2)	Accuracy (%)					
	K = 3		K = 5		K = 7	
	Org	Mod	Org	Mod	Org	Mod
21-79	96.83	93.67	97.01	93.84	97.18	94.02
36-64	96.83	93.49	97.01	94.02	97.18	93.67
51-49	96.83	94.02	97.01	94.37	97.18	94.02
82-18	96.83	93.14	97.01	93.14	97.18	92.79
18-82**	96.83	92.44	96.83	92.79	97.18	92.44
44-56**	96.66	93.32	96.83	91.56	97.01	92.09
60-40**	97.01	75.08	96.49	75.78	97.19	74.73
74-36**	96.48	93.14	96.83	93.67	96.83	93.67

**Skip 135 records for data distribution

Table 10 Privacy measures on IRIS

Method	VD	CP	RP	CK	RK
WHT+WHT	1.2737	0.50	50.9770	0.50	0.7623
WHT+DCT	0.9345	0.5000	53.2433	0.500	0.0067
WHT+FISIP	0.6704	0.5	54.1233	0.5	0.0050
WHT+SAN	0.7944	0	48.1533	1	0.1033

Table 11 Privacy measures on WDBC

Method	VD	CP	RP	CK	RK
WHT+WHT	5.0548	9.8000	174.4946	0	0.0062
WHT+DCT	4.2033	9.7333	185.1981	0.0333	0.0019
WHT+FISIP	3.1417	10.8667	173.1258	0	0.0024
WHT+SAN	0.9609	9.3333	180.2974	0.0333	0.0021

6 Conclusion and Future Work

This paper proposes a new approach for Horizontal PPDM based on combination of linear transformations. It is a simple and efficient approach to protect privacy of individual customers by inference from the experimental results. This approach will be extended in future to more than two clients and different combinations of transformation techniques and will be applied to vertically partitioned data also.

References

1. J.F. Traub, Y. Yemini, and H. Wozniakowski, "The StatisticalSecurity of a Statistical Database," ACM Trans. Database Systems, vol. 9, no. 4, pp. 672–679, 1984.
2. C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Ninth Int'l Conf. Extending Database Technology, pp. 183–199, 2004.
3. D. Defays and P. Nanopoulos, "Panels of Enterprises andConfidentiality: The Small Aggregates Method," Proc. Statistics Canada Symp. 92 Design and Analysis of Longitudinal Surveys, pp. 195–204, 1993.
4. J. Domingo-Ferrer and J.M. Mateo-Sanz, "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," IEEE Trans. Knowledge and Data Eng., vol. 14, no. 1, pp. 189–201, 2002.
5. C.C. Yao, "How to generate and Exchange Secrets", IEEE, 1986.
6. M. Kantarcioglu and C. Clifton. "Privately computing a distributed k-nn classifier". PKDD, v. 3202, LNCS, pp. 279–290, 2004.
7. Z. Yang, S. Zhong, R. Wright, "Privacy-preserving Classification of Customer Data without Loss of Accuracy", In: Proceedings of the Fifth SIAM International Conference on Data Mining, pp. 92–102, NewportBeach, CA, April 21–23, 2005.
8. L. Xiong, S. Chitti and L. Liu. k Nearest Neighbor Classification across Multiple Private Databases. CIKM'06, pp. 840–841, Arlington, Virginia, USA, November 5–11, 2006.
9. M. Kantarcioglu and J. Vaidya. Privacy preserving naïve Bayes classifierfor horizontally partitioned data. In IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, pp. 3–9, November 2003.
10. ZhuojiaXu, Xun Yi, "Classification of Privacy-preserving Distributed Data Mining Protocols", IEEE, 2011.
11. Jen-Wei Huang, Jun-Wei Su and Ming-Syan Chen, "FISIP: A Distance and Correlation Preserving Transformation for Privacy Preserving Data Mining" IEEE, 2011.
12. Shibnath Mukharjee, Zhiyuan Chen, Aryya Gangopadhyay, "A Privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms", the VLDB Journal, pp (293–315), 2006.
13. http://kdd.ics.uci.edu/.
14. http://www.wekaito.ac.nz/ml/weka.
15. ShutingXu, Jun Zhang, Dianwei Han, and Jie Wang, (2005) "Data distortion for privacy protection in a terrorist Analysis system", P. Kantor et al (Eds.): ISI 2005, LNCS 3495, pp. 459–464.