Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Transformation and Encryption Based Privacy Preserving Data Mining System

Santosh Kumar Bhandare
Asst. Professor
Department of Computer Science and Engineering
Shri Dadaji Institute of Technology & Science
Khandwa (M. P.) 45000, India

Abstract*— Data mining system contain large amount of private and confidential data. Privacy preserving data mining system is a popular and important research area. Privacy preservation of data in data mining system is required to avoid privacy leakage. We need an algorithm which protects private and sensitive information from large amount of data in data mining system. The data perturbation and cryptographic technique is one of the best methods for preserving privacy. In this paper we proposed an efficient privacy preserving data mining technique based on data perturbation combine with cryptographic technique. In this technique we perform both data transformation and encryption technique on some or all the information before applying data mining application. We are interesting to apply this algorithm on real life datasets.

Keywords*— Data mining, Normalization, Cryptography, Asymmetric encryption, privacy preservation.

I. INTRODUCTION

Data mining [1] is the process of finding pattern from large amount of data using tool such as classification. The problem of privacy preserving is very important issue in data mining system. There are a lot of data mining application deal with privacy and security concern. Data mining system contain large amount of private and secure data i.e. financial, criminal and healthcare records. These records cannot be share to everyone so privacy of data is required for avoiding privacy leakage. There are a lot of research has been done in privacy preserving data mining system. In this paper we discuss the combine strategy for privacy preserving data mining based on data perturbation and cryptography technique in which confidential numerical attributes are first distorted by data transformation method then the distorted datasets are encrypted by cryptography technique for privacy protection of data. Cryptography technique is one of the best techniques for privacy preserving data mining system. In the proposed technique the individual data values are transformed and then encrypted before applying data mining application. In this paper we proposed combination of any data transformation method and any cryptography technique for preserving private and confidential data. We are interesting to apply the proposed technique on real life dataset for privacy protection.

II. RELATED WORK

There has been a lot of privacy preserving data mining literatures. These literatures can divide into two categories. In the first category, methods modify the data mining algorithms so that without knowing the exact values of data, they allow data mining operations on distributed dataset. In the second category, methods are modifying the values of the datasets to protect privacy of data values. In this category there are several research has been done in data distortion or data perturbation are as follow: In the year 1985, Liew et al., they proposed data distortion method based on probability distribution [2]. This method involves three steps: (i) identification of the underlying density function, (ii) generation of a distorted series from the density function, and (iii) mapping of the distorted series onto the original series.

In the year 2000, Agrawal R. et al.,[3] they proposed an additive data perturbation method for building decision tree classifiers. Every data element is randomized by adding some noise. These random noise chosen independently by a known distribution like Gaussian distribution. The data miner rebuilds the distribution of the original data from its distorted version. In the year 2002, Sweeney L. et al., [4] in this paper the k-Anonymity model consider the problem that a data owner wants to share a collection of person-specific data without revealing the identity of an individual. This goal is achieve by data generalization and suppression methods are used to protect the confidential information.

This paper also examines the re-identification attacks. In the year 2005, Chen et al., they proposed a rotation based perturbation method [5]. The proposed method maintains zero loss of accuracy for many classifiers. Experimental results show that the rotation perturbation can greatly improve the privacy quality without sacrificing accuracy. In the year 2006, Wang et al., has used the Non-negative matrix factorization (NNMF) for data mining [6]. In this work, they combined non-negative matrix decomposition with distortion processing. The presented method have two important aspects (i) non-negative matrix factorization (NMF) is used to provide a least square compression version of original datasets and (ii) Using iterative methods to solve the least square optimization problem is provided an attractive flexibility for data administrator. The presented result given that the careful choice of iterative parameter settings, two sparse non-negative factors can solve by some efficient algorithms. Alternating least square using projected gradients in computing NNMF converges faster than multiplicative update methods. Iterative NMF based distortion method provides good solution for data mining problem on the basis of discriminate functions. In the year 2006, Li et al., proposed kd-tree based [7]. This method recursively partitions a data set into smaller subsets, so that the data records within each subset are more homogeneous after each partition. Then the confidential data in each final subset are perturbed using the subset average. This method is both efficient and effective, due to the recursive divide-and-conquer technique used. The experimental results show that the proposed method is effective.

In the year 2006, Xu et al., proposed Singular value decomposition (SVD) based data distortion strategy for privacy protection [8]. In this work they propose a sparsified Singular Value Decomposition (SVD) method for data distortion. They conducted experiment on synthetic and perturbation Method real world datasets and the experimental result show that the sparsified SVD method is effective in preserving privacy as well as maintaining the performance of the datasets.

In the year 2006, Wang et al., they proposed a new data distortion method based on Structural Partition and SSVD for Privacy Preservation [9]. They used object-based partition, feature-based partition and hybrid partition. The experimental results show that feature-based partition is a feasible and efficient solution for privacy-preserving data mining.

In the year 2007, Saif et al., also used non-negative matrix factorization for data perturbation [10]. They investigated the use of truncated non-negative matrix factorization (NMF) with sparseness constraints. The experimental results show that the Non-negative matrix factorization with sparseness constraints provides an efficient data perturbation tool for privacy preserving data mining. The privacy parameter used in the proposed work provides some indication on the ability of these techniques for concealing the original data values.

In the year 2007, Wang et al., they proposed several efficient and flexible techniques to address accuracy issue, in privacy preserving data mining through matrix factorization [11]. Experimental results indicate that for centralized datasets with numerical attributes, matrix factorization-based distortion strategies achieve a satisfactory performance.

In the year 2007, Xu et al., has used the Fast Fourier Transform (FFT) for data perturbation [12]. The dataset is distorted or perturbed by using Fast Fourier Transform (FFT) for privacy protection of data values.

In the year 2008, Liu et al., has used the wavelet transformation for data distortion or data perturbation to preserve the privacy of data [13]. Privacy preserving strategy based on wavelet perturbation; keep the data privacy and data statistical properties and data mining utilities at the same time. The results show that presented method keep the distance before and after data perturbation and it also preserve the basic statistical properties of original data while maximizing the data utilities.

In the year 2009, Lin et al., has presented a method for data perturbation. In this method, the data matrix is vertically partitioned into several sub-metrics and held by different owners [14]. For perturbing their individual data, each data holder can randomly and independently choose a rotation matrix. The presented results show that random rotation based method for data perturbation preserve the data privacy without affecting the accuracy.

In the year 2010, Peng et al., used combine data distortion strategies for privacy preserving data mining [15]. They designed four schemes via attribute partition, with single value decomposition (SVD), non-negative matrix factorization (NMF), discrete wavelet transformation (DWT) for distortion of submatrix of the original dataset for privacy preserving. The basic idea of the proposed strategies was to perform distortion on sub matrices of original dataset using different method. The results show that proposed method was very efficient in maintaining data privacy as well as data utility in comparison to the individual data distortion techniques such as SVD, NMF and DWT.

In the year 2011, Keng-Pei Lin et al., [19] they propose an approach to post process the SVM classifier to transform it to a privacy-preserving classifier which does not disclose the private content of support vectors. The post processed SVM classifier without exposing the private content of training data is called Privacy-Preserving SVM Classifier (abbreviated as PPSVC). The PPSVC is designed for the commonly used Gaussian kernel function. It precisely approximates the Decision function of the Gaussian kernel SVM classifier without exposing the sensitive attribute values possessed by Support vectors. By applying the PPSVC, the SVM classifier is able to be publicly released while preserving privacy. We prove that the PPSVC is robust against adversarial attacks.

The experiments on real data sets show that the classification accuracy of the PPSVC is comparable to the original SVM classifier.

III. DATASET

The object-attribute relationship of real life data sets are encode into vector – space format [16]. In this format a 2-dimentional is used to share the dataset. Row of the matrix indicates individual object and each column represent a particular attribute of these objects. In this matrix, we assume that every element is fixed, discrete and numerical. Any missing element is not allowed

We are interesting to apply our proposed privacy preserving technique on real life datasets. We will choose four real-life Databases obtained from the University of California Irvine (UCI), Machine Learning Repository [17]. Datasets are the Glass Identification, Haberman's survival data, Bupa Liver Disorders and Iris Dataset. The summaries of the original database are given in Table [I]. We will use WEKA (Waikato Environment for Knowledge Analysis) [18] software to test the accuracy of datasets.

TABLE I: The summary of the database

Database	Number of Instances	Number of Features	Number of Classes
Glass Identification	214	10	7
Haberman's Survival data	306	3	2
Bupa Liver Disorders	345	6	2
Iris	150	4	3

IV. COMBINE STRATEGY FOR PRIVACY PRESERVING DATA MINING

In the combine strategy for privacy preserving data mining based on data transformation and encryption we must first transformed the original dataset by any data transformation method for obtaining distorted datasets now this transformed datasets will be encrypted by any encryption technique. After applying the encryption algorithm we will obtain the strong distorted and encrypted version of original dataset. This distorted and encrypted datasets is now sending to data miner for applying data mining algorithm. The data miner obtains the original dataset by applying the reverse process on the distorted and encrypted datasets and apply data mining algorithm on it. The result of the data mining algorithm is now sending to the datasets owner by the same process.

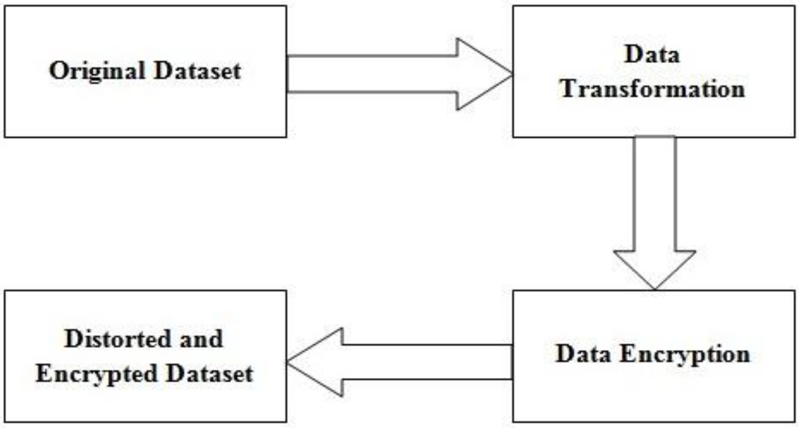

Fig 1. Privacy Preserving Methodology

V. CONCLUSION

In data mining system there are so many privacy preserving method. In this paper we proposed an efficient privacy preserving algorithm for protecting confidential and private data in data mining system. In this technique we perform both data transformation and encryption technique on some or all the information before applying data mining application. This is the strong privacy preserving data mining technique because we provide two level of security in our proposed method first is data transformation and second is data encryption. Therefore it is the more secure privacy preserving data mining technique. We are interesting to apply our proposed technique on real life dataset and compare its result with the existing privacy preserving method.

REFERENCES

[1] M. Chen, J. Han, and P. Yu, "Data mining: An Overview from a database Prospective", IEEE Trans. on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996.

[2] C. K. Liew, U. J. Choi, and C. J. Liew, "A Data Distortion by Probability Distribution", ACM Transaction on Database Systems (TODS), vol. 10, no. 3, pp. 395-411, Sep. 1985.

[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining", Proceeding of the ACM SIGMOD Conference on Management of Data, pp. 439–450, May 2000.

[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[5] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation", Proceeding of the 5th IEEE International Conference on Data Mining (ICDM 2005), pp. 589-592, 2005.

[6] Jie Wang, Weijun Zhong, Jun Zhang, "NNMF- Based Factorization Techniques for High-Accuracy Privacy Protection on Non-negative-valued Dataset", Proceeding of IEEE Conference on Data Mining, International Workshop on Privacy Aspects of Date Mining (PADM2006), pp.513-517, 2006.

[7] Xiao-Bai Li, and Sumit Sarkar, "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 9, pp. 1278 – 1283, 2006.

[8] S. Xu, J. Zhang, D. Han and J. Wang, "Singular value decomposition based data distortion strategy for privacy protection", ACM Journal of Knowledge and Information Systems, vol. 10, no. 3, pp. 383-397, 2006.

[9] J. Wang, W. J. Zhong, J. Zhang and S.T. Xu, "Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation", Proceedings of International conference on Information & Knowledge Engineering, pp. 114-120, June 2006.

[10] Saif M. A. Kabir, Amr M. Youssef, Ahmed K. Elhakeem, "On data distortion for privacy preserving data mining", Proceedings of IEEE Conference on Electrical and Computer Engineering (CCECE 2007), PP. 308-311, 2007.

[11] Jie Wang, Jun Zhang," Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization", pp. 217-220, 2007.

[12] Shuting Xu, Shuhua Lai,"Fast Fourier transform based data perturbation method for privacy protection", Proceeding of IEEE International Conference on Intelligence and Security Informatics, pp. 221-224, 2007.

[13] Lian Liu, Jie Wang, Jun Zhang,"Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving", Proceeding of IEEE International Conference on Data Mining Workshop, PP. 27-35, 2008.

[14] Zhenmin Lin, Jie Wang, Lian Liu, Jun Zhang, "Generalized random rotation perturbation for vertically partitioned data sets", Proceeding of the IEEE Symposium on Computational Intelligence and Data Mining, pp:159- 162, 2009.

[15] Bo Peng, Xingyu Geng, Jun Zhang, "Combined data distortion strategies for privacy-preserving data mining", Proceeding of the IEEE International Conference on Advanced Computer Theory and Engineering (1CACTE), PP. V1-572 - V1-576, 2010.

[16] W. Frankes and R. Baeza-Yates, "Information Retrieval: Data Structures and Algorithms", Prentice–Hall, Englewood cliffs, NJ, 1992.

[17] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html

[18] The Weka Machine Learning Workbench. http://www.cs.waikato.ac.nz/ml/weka

[19] Keng-Pei Lin and Ming-Syan Chen, "On the Design and Analysis of the Privacy-Preserving SVM Classifier", Proceeding of the IEEE Transactions on Knowledge and Data Engineering, VOL. 23, NO. 11, pp. 1704-1717, NOV 2011

AUTHOR

Mr. Santosh Kumar Bhandare is presently working as Asst. Professor in Department of Computer Science & Information Technology at Shri Dadaji Institute of Technology & Science Khandwa (M.P.) 450001 India. The degree of B.E. secured in Computer Science & Engineering from Madhav Institute of Technology & Science Gwalior in 2006, M.Tech in Computer Science & Engineering from Samrat Ashok Technological Institute Vidisha in 2011. Research Interest includes Data Mining, Network Security, Computer Graphics & Multimedia, Privacy Preserving Data Mining, and Cloud Computing.

Mobile: +91-9827099815, **E-mail**: santosh.mits@gmail.com