www.ijiarec.com

ISSN: 2348-2079

MAR-2015

International Journal of Intellectual Advancements and Research in Engineering Computations

DOUBLE KEY SEEDS BASED PRIVACY PRESERVING DATA MINING APPROACH FOR ENHANCED REVERSIBLE DATA HIDING

[1]R.Kavitha, [2]D.Vanathi, [3]Dr.P.Sengottuvelan

ABSTRACT

The ultimate goal of Privacy Preserving Data Mining is to develop efficient algorithms that allow one to extract relevant knowledge from large amount of data, while preventing sensitive information from disclosure or inference. Several data mining algorithms, incorporating privacy preserving mechanisms, have been developed over the last years. There is an emerging need of synthesizing literature to understand the nature of problem and evaluate the relative performance of different approaches. In the proposed privacy optimal aggregation (POA) method, the original data is perturbed and embedded with a fragile watermark to accomplish privacy preserving and data integrity of mined data and to recover the original data. This technique is a high quality reversible method with crypto analysis for data embedding. A Special case of information hiding is digital watermarking poses double key distribution for data recovery process. The proposed method is for protecting privacy data as well as preventing illegal users from searching correlation of the distributed data security.

KEYWORDS: privacy, watermarking, cryptography, heuristic approach, Privacy Optimal Aggregation

INTRODUCTION

Data mining is the process of extracting information about the user data, also called knowledge invention on internet. With the advent of the Internet and new technologies that allow easier, quicker, as well as anonymous access to more information than ever before, people have now become more aware of identity theft and make conscious decisions on how to protect themselves. Effortless access to such private data causes a risk to individual privacy. Official statistics, Health information, and E-commerce are some key concern for privacy. Privacy preserving data mining technique gives novel way to solve this problem. The main purpose of privacy preserving data mining is to design proficient frameworks and algorithms that can extract relevant knowledge from a large amount of data without revealing of any sensitive information. It protects sensitive information by providing sanitized database of original database on the internet or a process is used in such a way that confidential data and private knowledge remain private even after the mining process. It is PPDM due to which the benefits of data mining be enjoyed, without compromising the privacy of concerned individuals.

The PPDM Techniques can be classified over five dimensions. The first aspect is related to distribution of data i.e. Centralized or Distributed. The second aspect refers to the modification of original values of data that are to be released for data mining task. Modification is carried out using perturbation, blocking, aggregation, merging, exchange or sampling or any grouping of these. The third aspect is that of data mining algorithms. The data mining algorithm are applied on the transformed data to get useful nuggets of information that were secret previously. The fourth aspect refers to whether the raw data or aggregated data should be secret.

The fifth and the final aspect refer to the techniques that are used for protecting privacy. Based

Author for Correspondence:

[1]PG scholar, Dept. of CSE, Nandha Engineering College, Erode, Tamilnadu, India. Email:Kavithamohan926@gmail.com

[2]Associate Professor, Dept. of CSE, Nandha Engineering College, Erode, Tamilnadu, India. Email:vanathi.d@nandhaengg.org

[3]Associate Professor, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India. Email:sengottuvelan@rediffmail.com

on these dimensions, different PPDM techniques may be classified into following categories.

A. Anonymization based PPDM, when quasi identifiers [set of attributes that could potentially recognize a record] are linked to publicly available data, identity of individual can be predicted with higher probability. Such attacks are called as linking attacks. Anonymization approach conceal identity or/and sensitive data about record owners using generality and suppression in anonymzed dataset. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Such data when released for mining reduces the risk of identification when combined with publicly available data. But besides accuracy of the applications on the transformed data is reduced.

B. Randomized Response based PPDM Randomized response is statistical technique to solve a survey problem. In Randomized response, the data is jumbled in such a way that the central place cannot tell with probabilities better than a pre-defined threshold, whether the data from a customer contains truthful information or false information. The information received from each individual user is snarled and if the number of users is significantly large, the aggregate information of these users can be predictable with good amount of accuracy. The data collection process in randomization method is carried out using two steps. During first step, the data providers randomize their data and transmit the randomized data to the data receiver. In second step, the data receiver reconstructs the original distribution of the data by employing a distribution reconstruction algorithm.

The randomization method can be implemented at data collection time. It does not need a trusted server to contain all the original records in order to perform the anonymization process. The weakness of a randomization response based PPDM technique is that it treats all the records equal irrespective of their local density. This leads to a problem where the outlier records become more prone to adversarial attacks than to records in more dense regions in the data. One solution to this is to unnecessarily adding noise to all the records in the data. But, it reduces the utility of the data for mining purposes as the reconstructed distribution may not yield results in consistency of the purpose of data mining.

HEURISTIC APPROACH

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1's to 0's and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns to prevent disclosure and privacy. These approaches have been getting attention for majority of the researchers due to their efficiency, scalability and quick responses.

For most image data hiding methods, the host image is permanently distorted and it cannot be restored from the marked content. A method called "Reversible Data Hiding" (RDH) is proposed, in which the host image can be fully restored after data embedding. RDH is a hybrid method which combines various techniques to ensure the reversibility. Its feasibility is mainly due to the lossless compressibility of natural images.

Many RDH methods have been proposed in recent years, e.g., the methods based on lossless compression, difference expansion (DE), histogram shifting (HS), and integer transform, etc. All these methods aim at increasing the embedding capacity (EC) as high as possible while keeping the distortion low.

The Proposed a new HS-based method by modifying the prediction-error histogram. This method can well exploit the image redundancy and thus achieve a better performance compared with the previously introduced DE-based methods. Recently, privacy and security have to propose a novel HS-based method where the histogram bins used for expansion embedding are specifically selected such that the embedding distortion is minimized. The experimental results reported demonstrated that this method is better than some state-of-the-art works including the most recently proposed integer-transform-based method.

The main objectives of the proposed system is,

1. The maximum modification to Data values can be controlled and thus the embedding distortion can be well limited.

2. The location map used to record underflow/overflow locations is usually small in size especially for low ER (Embedding Rate) case. It has more accuracy.

PROBLEM FORMULATION

In previous, the privacy preserving data mining (PPDM) can prevent private data from discovery in data mining. However, the current PPDM methods damaged the values of original data where knowledge from the mined data cannot be verified from the original data. We combine the concept and technique based on the reversible data hiding the reversible privacy preserving data mining scheme in order to solve the irrecoverable problem of PPDM. In this method, the original data is perturbed and embedded with a fragile watermark to accomplish privacy preserving and data integrity of mined data and to also recover the original data. Experimental tests are performed on classification accuracy which is used to evaluate the efficiency of POA for privacy preserving and knowledge verification. However, this method and its extensions all encompass the weak spot of a built-in threshold determined by the measure of the polynomial: when the numeral of messages transmitted is better than this threshold, the challenger can completely recover the polynomial the theoretical examination and simulation results display that scheme is more efficient than the polynomial-based move toward in terms of computational and message overhead under comparable safety levels while as long as message foundation time alone.

PROPOSED TECHNIQUES

The heuristic approach in data mining is to make the correlation of encrypted data hiding in privacy is more efficient to enhance security against random timing perturbations by the adversary model. Compared to existing timing-based correlation schemes, our watermark-based correlation in data mining is active in that it embeds a unique watermark into the encrypted flows of data hiding, by slightly adjusting the knowledge discovery and reversible data hiding algorithm of selected datasets for hiding the data in privacy secured level .If the embedded watermark is both unique and robust, the watermarked flows can be effectively identified and thus correlated at each stepping stone can be retrieval's more efficient. Watermark into the image (the spatial domain) by interpolating the intensity value of the original pixels in the image. These spatial domain-based watermarking embedding techniques can embed relatively large amounts of data into the image. However, they generally are not robust to image distortions. Consequently, recent watermarking techniques do not directly change the pixel values in the image. Instead, they first transform the image into another frequency domain by applying any of several methods like difference expansion, histogram based method. In this Efficient Heuristic approach method is proposed which is more effective to hide association rule. The objective of this algorithm is to extract relevant knowledge from privacy of data .In Special case information privacy requires randomized key distribution poses double key verification i.e., reversible key ordering and distribution security key While protecting at the time sensitive information. The proposed method focused on hiding set of frequent items containing highly sensitive knowledge that only remove information from transactional database with no hiding failure.

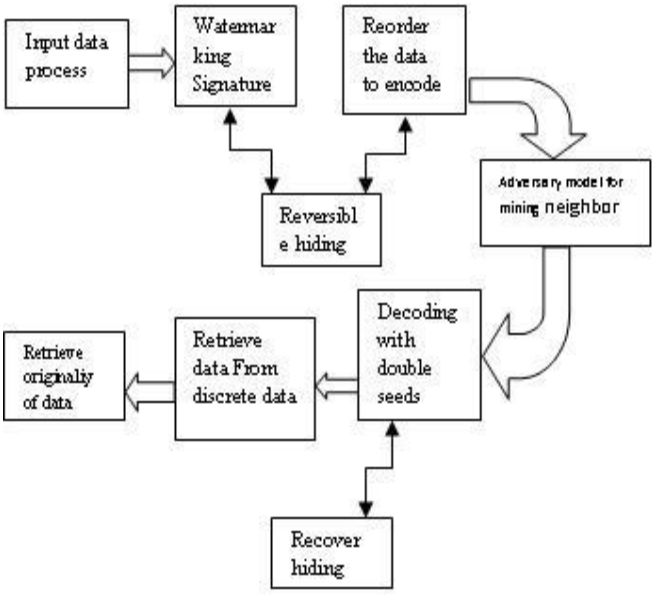

Figure 3.1. Proposed block diagram

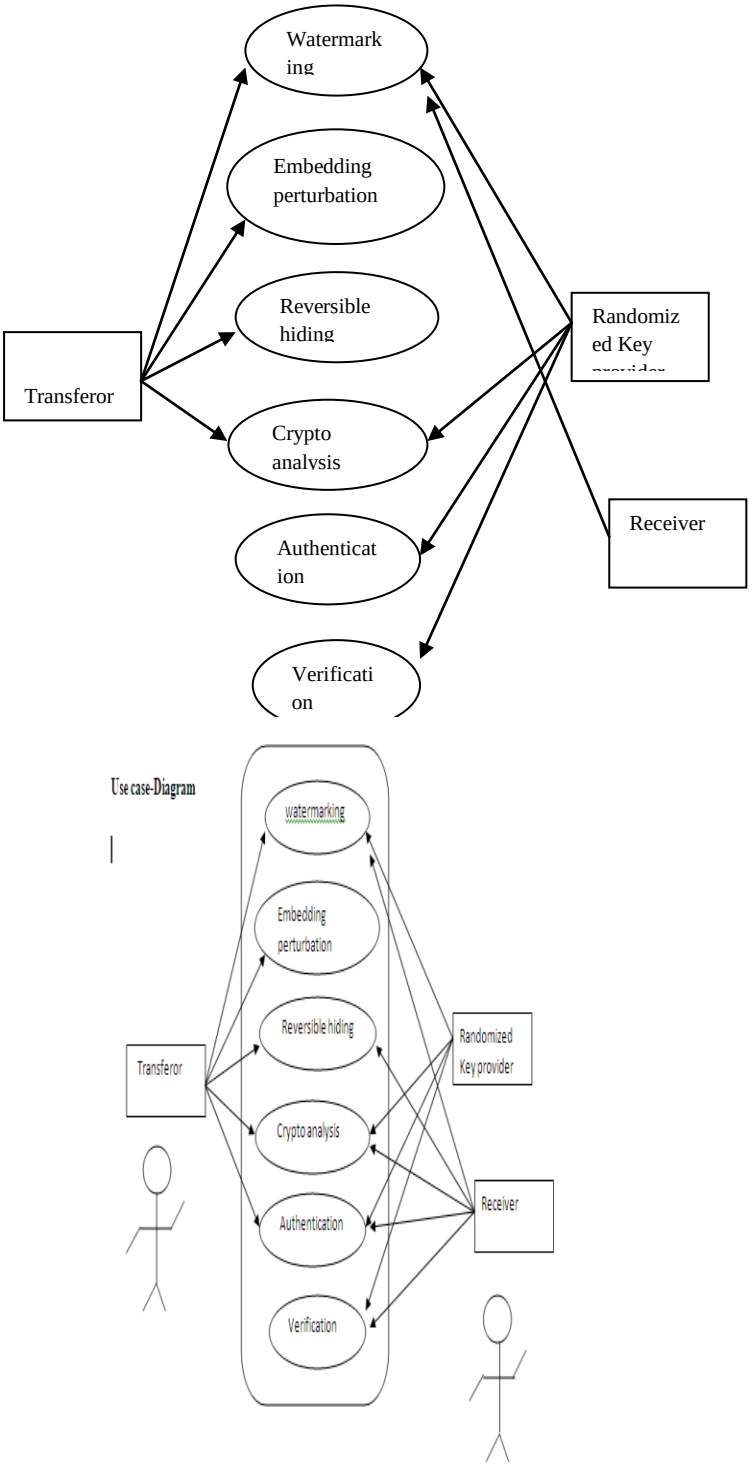

Figure 3.2 Proposed methods – use case diagram.

Figure 3.1 shows the proposed novel system architecture. Figure 3.2 shows the use case diagram .

The aim of this paper is to provide a sufficient method for recovery of the original data and to protect the data's privacy.

Difference expansion method is used to perturb data before embedding into image which is illustrated below.

POA OPTIMAL AGGREGATION ALGORITHM

STEPS OF THE PROPOSED ALGORITHM

Original data *D*, where *D* = {*di*, *i* = 1, 2, 3, . . . , *n*}, *n* is the data number of *D*, *m* is the number of perturbed attributes, *s* is the watermark, and *w* is the bit length of *s*.

PERTURBED PHASE ALGORITHM

D^2 is the perturbed mining data, D^2 = {$_{d'i}$, *i* = 1, 2, 3, . . . , *n*}. The key of the secure permutation which is the *Seed* value.

Algorithm: Privacy Optimal Aggregation, called as POA(Eps, k)

Input: A dataset D of n records and the value k for k-partition.

Step1: Partition dataset D, with optimal SCAN (Eps, k) as C ={ C1, C2, □ , Cn}, such that CiCj = Φ, i ≠ j;N:= D–i

Step 2: Embed the watermark *sx* into to nearest cluster using k-nearest-neighbour tree aggregation;

Step 3: For each, Ci □ C, |Ci=1,n| ≥ 2k, Call matching (Ci) to partition Ci := {Pi1, Pi2, □ , Pin, □ k ≤ |Pi=1,n| ≤ 2k-1}; 4. End;

Step 4: Reorder the partitioned data in *D* by a secure permutation.

Before recovering the original mining data *D*, we must first reorder the permutated *D* and recover the neighboring data groups. Output updates D, as the transformed D.

For example, we must use the *Randomize (Seed)* × *n* function to generate the same series, re-sort *D* data records according to this series, and then recover the original neighboring data groups.

RECOVERY PHASE

The recovery phase algorithm is listed as follows.

The permutated mining data *D* , *D* = {*d* ‘ *i*, *i* = 1, 2, 3, . . . , *n*} where *n* is the data number of *D* , the number of attributes *m*, perturbation parameter θ , and the key of the secure permutation.

ALGORITHM: Original data *D*, *D* = {*di*, *i* = 1, 2, 3, . . . , *n*}, watermark *s*, *w* is the bit length of *s* .

Step 1: Reordering the permutated D_2 and recovers the neighboring data groups.

Step 2: Calculate the hidden watermark *s* , and the aggregate value *diff* of the original neighboring data group. Repartioning the data group is done.

Step 3: Recover the original neighboring data group and recover the original data *D*.

Step 4: If *j* ≤ *m*, then *j* = *j* + 1 and go back to *Step2*.

In recovery process of algorithm, restores the original data *D*, and the original data is verified by the watermark *s*. (check for *s* x equals to *s*, otherwise the perturbed data *D* is distorted).

POA ALGORITHM EVALUATION USING WEKA TOOL

Figure 5.1 shows the flow diagram of the algorithm evaluation that uses WEKA tool. Experiments are set up to evaluate the performance of data perturbation method. Waikato Environment for Knowledge Analysis (WEKA) tool is used to test the accuracy of POA algorithm. The data perturbation algorithm is implemented by a separate Java programme.

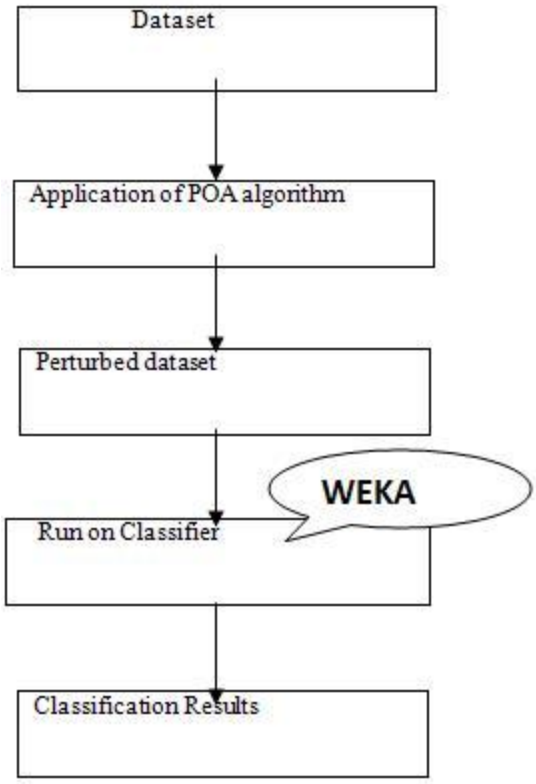

Figure 5.1. Flow Diagram for Evaluation of POA

Following are the basic steps for how to perform whole experiment.

Step 1. Generate a dataset or take a dataset. Here a dataset is taken from UCI data repository.

Step 2. Apply the POA algorithm on dataset and generate perturbed dataset.

Step 3. Take one classification algorithm (FLR or Naïve Baise) and apply on perturbed dataset, by using WEKA tool .

Step 4. Analyse classification results

RESULTS

Two datasets Vehicle and Abalone Dataset are taken from UCI dataset repository. The numbers of attributes for these two datasets are 19 and 8 respectively.

Table 6.1 The test datasets

Datasets	**Number Of Attributes**	**Number Instances**	**Number Of Classes**
Abalone	8	4177	3
Vehicle	19	846	4

According to the classification accuracy, the results are similar before and after by POA in Abalone and Vehicle datasets. It means that data perturbing by POA for these

two datasets can reserve the correct knowledge.

The experimental results show that POA provides privacy protection for PPDM. Also, POA can restore mining data back to the original, and the results of knowledge analysis are similar to the original data. POA not only ensures the value of knowledge analysis of original data,but also protects the privacy data.

CONCLUSIONS

An efficient double seeds key distribution for privacy preserving using heuristic based data embedding method. The attributes are scanned as an embedding unit to preserve the data and a specially designed neighborhood set is employed to embed message digits with a smallest notational system. It allows users to select digits in any notational system for data embedding, and thus achieves a better image quality. The proposed method enhances the double seed security Moreover, it produces no artifacts false points security in watermarked images and the watermarked analysis results are similar to those of the cover images, it offers a secure communication under adjustable embedding capacity.

This method combines RDH and PPDM techniques to perform privacy preserving. The privacy data is perturbed by the aggregate value of clustered neighboring data, which we called POA. POA efficiently balanced data protection and knowledge mining. One aim is to achieve privacy protection by PPDM, while another is to allow legal users to

restore private data, and the relationship is further analyzed between the original data and knowledge.

Second part, the classification accuracy (USING Weka tool) is same for perturbed dataset. There is scope for further improvement in proposed methods and algorithms for preserving privacy. Output of such methods can be compared based on widely accepted evaluation metrics. Further we can also propose evaluation metrics to measure information gain / loss and privacy gain.

REFERENCES

[1] Aggarwal CC, Yu PS (2008) Privacy-preserving data mining: models and algorithms. Springer, Berlin

[2] Agrawal R, Srikant R (2000) Privacy-preserving data mining. SIGMOD Rec. 29(2):439–450. doi:10.1145/335191.335438

[3] 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, SEPTEMBER 2006

[4] Murat Kantarcioglu, Bhavani Thuraisingham ,"The applicability of the perturbation based privacy preserving data mining for real-world data", IEEE, Li Liu, 2007

[5] Hillol Kargupta , Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar ,"On the Privacy Preserving Properties of Random Data Perturbation Techniques",ICDM2003.Third IEEE International conference on 19-22Nov 2003.

[6] Yingpeng Sang, Hong Shen, and Hui Tian, "Effective reconstruction of data perturbed by random projections", IEEE TRANSACTIONS ON COMPUTERS, VOL. 61, NO. 1, JANUARY 2012

[7] Kun Liu, Hillol Kargupta, Senior Member, and Jessica Ryan ," Random projection based multiplicative data perturbation for privacy preserving distributed data mining ",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 1, JANUARY 2006

[8] Amari S, Wu S (1999) Improving support vector machine classifiers by modifying kernel functions. Neural Netw 12(6):783–789. doi:10.1016/S0893-6080(99)00032-5

[9] Census Bureau US (2011) Census bureau homepage. http://www.census.gov/. Accessed 20 June 2011

[10] Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16(1):27–31. doi:10.2174/092986609787049420

[11] Chen TS, Chen J, Lin YC, Tsai YC (2009) Research to protect database by shaking random sampling interference (SRSI). In: Proceedings of the 2009 global congress on intelligent systems, pp 569–572. doi:10.1109/GCIS.2009.384

[12] Chun JY, Hong D, Jeong IR, Lee DH (2011) Privacy-preserving disjunctive normal form operations on distributed sets. Bibliogr. - Inst. Presse Sci. Inf. 231:113–122

[13] Domingo-Ferrer J, Mateo-Sanz JM, Torra V (2001) Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Proceedings of the international conference on new techniques and technologies for statistics: exchange of technology and knowhow, pp 807–826

[14] Frank A, Asuncion A (2010) UCI machine learning repository. http://archive.ics.uci.edu/ml/. Accessed 1 March 2011

[15] Furey TS, Cristianini N, Duffy N, Bednarski DW (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914. doi:10.1093/bioinformatics/16.10.906