International Journal of Engineering & Technology, 7 (2.7) (2018) 515-519

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



A Novel Privacy Preserving Data mining using improved decision tree and KP-ABE on High Dimensional Data

Aaluri Seenu [1], M. Kameswara Rao [2*]

[1] Research Scholar, Department of CSE, KLEF
[2] Associate Professor, Dept of Electronics and Computer Engg , KLEF
**Email: kamesh.manchiraju@kluniversity.in*

Abstract

In distributed data mining environment maintaining individual data or patterns is a major issue due to high dimensionality and data size. Distributed Data mining framework can help to find the essential decision making patterns from distributed data. Privacy preserving data mining (PPDM) has emerged as a main research area for data confidentiality and knowledge sharing in between the communicating parties. As the distributed data of the individuals are stored by the third party, it leads to the misuse of distributed information in digital networks. Most of the decision patterns generated using the machine learning models for business organizations, industries and individuals has to be encoded before it is publicly shared or published. As the amount of data collected from different sources are increasing exponentially, the time taken to preserve the patterns using the traditional privacy preserving data mining models also increasing due to high computational attribute selection measures and noise in the distributed data. Also, filling sparse values using the conventional models are inefficient and infeasible for privacy preserving models. In this paper, a novel privacy preserving based classification model was designed and implemented on large datasets. In this model, a filter-based privacy preserving model using improved decision tree classifier is implemented to preserve the decision patterns using IPPDM-KPABE model. Experimental results proved that the proposed model has high computational efficiency compared to the traditional privacy preserving model on high dimensional datasets.

Keywords: *Privacy Preserving Data Mining, Decision Trees, ABE.*

1. Introduction

Data anonymization is the process of replacing the contents of essential private patterns such as data attributes, ranking measures, decision patterns etc , so that records cannot be associated with a specific individual, project or company and yet to preserve a balance between privacy and data utility. Privacy-Preserving Data Mining (PPDM) integrates the concepts of data mining models with the probable privacy protection techniques [1]. PPDM is responsible for preserving extended security to the sensitive information to the data mining framework.

A large number of privacy preserving techniques have been implemented using the data mining model to provide hidden privacy on the large volumes of data [1]. It is very essential and important to maintain the privacy preservation in the distributed environment.The following are the various factors that can be adapted for implementing privacy preserving data mining(PPDM) models. 1. Distributed data distribution 2. data variations 3. data mining models 4. Hidden patterns 5. Privacy preservation. First factor refers to distributed data distribution from different sources; how data is distributed or located centralized or distributed manner. In distributed databases, data can be distributed horizontally or vertically for privacy preserving procedure. Second factor refers to variation of the original data to other pre-processed form to filter the sensitive data, here actual data will be transformed by multiplying the noise to some extent, Third factor is applying the data mining algorithm on input data to preserve privacy patterns of attribute values. Here, data mining is used to find the essential hidden knowledge patterns on huge volume of data. Knowledge extraction models includes classification , association ,clustering and prediction algorithms are implemented to find and extract the hidden and valid association rules, decision patterns or clusters from huge datasets. Fourth factor refers to pattern hiding, some part of input data or feature or output of the data mining will be taken and keep them in hidden state. Fifth factor is the most important issue i.e., providing privacy during data sharing or representation or data visualization process.

Replication and Duplication can work to maintain the data in a current state at all locations the data is distributed to. Besides using distributed database replication and the fragmentation of data, there are also other distributed database-design technologies that can be used. Some of the other technologies that can be used are synchronous and asynchronous distributed database and local autonomy technologies. The Development of replication and duplication, fragmentation depends on applications of organizations.

Privacy preserving data mining approach can be broadly classified into two sub-categories; those are mentioned below [1-2].

1. The initial phase of PPDM approach focuses on protection of sensitive information. Some examples of such sensitive information are:- name, ID, address and other private details.
2. After completion of the initial phase of PPDM, the second phase starts with knowledge hiding in database. It is responsible for adding security measure to protect the sensitive information. The prime objective of this phase is to determine the exact process of information hiding without influencing non-sensitive and useful rules.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Along with it, it is been observe and identify criteria to evaluate a PPDM algorithm:

1. Privacy level: This indicates how approximately the hidden sensitive information may still be estimated.
2. Hiding failure: This refers to that part of sensitive information which is not concealed by the employment of a privacy-preservation technique.
3. Data quality: This becomes important after a privacy-preserving technique has been applied. This looks at the quality of data as also the quality of the results derived from data mining after the concealing strategies were applied.
4. Complexity: This reveals how well a privacy-preserving algorithm can be executed considering the interaction of all the resources as may be implied by the algorithm.

The main limitations of the traditional privacy preserving models on distributed dataset are:

1. Information loss: When the original information is privacy preserved, it is assumed to be an information loss during the data recovery phase.
2. Privacy Preserved: Here, information is hidden or distorted. If the PPDM results higher value of privacy preservation, then the algorithm is considered to be a efficient one.
3. Computational time: Computational time plays an important role on deciding the efficiency of privacy preservation approach. A privacy preservation approach is efficient, if and only if it has more privacy with less computational time. In other words, decrease in computational time will enhance the efficiency and effectiveness of privacy preservation techniques.
4. Complexity: When the PPDM is applied on large datasets with minimal complexity, its efficiency and performance increases.
5. Dependency on size of data: Now-a-days with the exponential growth of data, the efficiency of algorithm decreases.

Privacy-preserving data mining has solutions that address these data privacy issues on large data size. As mentioned, data randomization is classified into two approaches. In the first approach, distributed data is to obscure or randomize the information before sending it to the data miner. In the second approach, privacy-preserving distributed data mining is applied. This method follows an approach where the data is portioned between two or more sites, and then these locations cooperate to discover the global data mining results without revealing any of the data they hold at their individual sites[3]

2. Related Work

Data perturbation approaches are used to provide privacy on sensitive data and to process data analysis. It is unable to reconstruct the original data, but only data distributions are reconstructed. Stanley R. M. Oliveira [5] proposed a privacy preserving clustering mining technique involving centralized distribution and rotational-based transformation. This approach was implemented on network dataset in which the data server is not permitted to learn or recover essential data records. This is one of the major limitations of perturbation approaches. All the traditional privacy preserving approaches are incapable of achieving extended privacy on large datasets. It is very important to add privacy measure prior to the distribution of data into multi-cloud servers. The most promising feature of every privacy preservation approach is, the customers' sensitive information is identified properly before sharing them with others. There are two key issues that need to be addressed in privacy preserving data mining, these are: Modification sensitive information such as attribute names, values and other group identifiers from the database to optimize the data privacy. The second issue is omission of sensitive knowledge that may be extracted using data mining procedures. Randomization refers to the technique through which input details are perturbed before being distributed to data mining algorithms in order to safeguard the data values of specific entities from being exposed. Privacy denotes the rights of users to hide private details about themselves and also to determine what may be disclosed.

Now-a-days, differential privacy preservation approach is most widely implemented in order to provide extended security to statistical databases. It is responsible for decreasing the chances of record identification for enhancement of privacy. On the contrary, the process of data security helps to protect data from unauthorized accesses while communicating through a network. When an authorized and validated user gets the data, there is no extra constraint enforced on that user from retrieving sensitive information. Therefore, many researchers have been tried to detect the correlation among data security and data privacy[4].

Privacy problems in data mining have to be handled with great care in data sharing. Several organizations collect information about customers' shopping behavior. To gain mutual benefit among themselves, they share data and conduct data mining on the databases for providing competitive advantages over others. The organizations while sharing the data with others must realize that it is necessary to shield sensitive data and patterns. The existing data sanitization methods proposed by [5] were capable of hiding sensitive patterns with limited side effects to the non-sensitive patterns. Yet, extended approaches are required to minimize the non-sensitive patterns affected. They also make it necessary to develop an ability to deal with different types of pattern. In the years past, several researchers have dealt with the problem of privacy preservation in frequent pattern mining. As far as the transactional databases are concerned, frequent pattern mining considers only the presence or absence of items. But in the real world, the transactional database represents the sale of each item. It is essential to consider the profit obtained in each transaction. [6] recommended a utility mining model to discover the itemsets, which yields high utility in the database. They emphasized a few sanitization approaches for privacy preservation of such itemsets. But those approaches affected a number of legitimate itemsets. In this thesis, two sanitization algorithms have been suggested to hide sensitive utility itemsets that further reduce the number of legitimate itemsets affected during sanitization. The first algorithm incorporates the conflict degree of a transaction to select the victim transaction. In the second algorithm, the conflict degree of an item is also used to identify the victim item for modification.

As the models do not have data values that are distinct the privacy threat is comparatively less here. There will be an inference issue. The flawless classifier will find the objective class, nevertheless the objective class is not revealed explicitly. However, in practical approach, the inferences which are probabilistic give the proprietor an assessment of individual values[7]. The data mining is not based on the results that are generated by instant privacy; however these methods are used to generate the results. In most of the cases, the statistical measures of data mining techniques such as association rules and classification rules are inefficient to apply privacy preservation. In this case, sometimes association rules are hidden to enhance the privacy level in data mining. A novel PPM known as Randomized response scheme [8] is developed to protect the respondent's privacy. It follows the basic concept of decomposition where the data owner divides original data and stores it at different places. It adds obstruction for the attacker to get the original data and sometimes some false information are also included to prevent attacks. [9] developed a new randomized response approach for performing different data mining calculations. This privacy preserving data mining approach is responsible for protecting respondent's privacy through random of questions. The proposed method is capable of determining appropriate distributed probability and also evaluates the distributed probability for certain characteristics. [10] developed a data hiding approach for privacy preservation association rule mining which depends on randomized response techniques. This technique protects the privacy of attributes which are in numerical format. This technique was prescribed to fulfil the requirements of classification analysis. Privacy preserving model using cryptographic approach describes

how various parties can calculate any function of their interest together and also ensure that no other information is revealed. As mentioned earlier, with the mentioned calculations, these results give utmost privacy where it provides privacy to the information except the function designed for the output[11][12-14].

3. Proposed Methodology

In this methodology, a novel privacy preserving based optimized decision tree classifier was implemented to address the problem of high dimensionality and privacy issues on large dataset. In this model, a novel decision tree classifier is developed to preserve the privacy among attributes using the PPDM-KPABE encryption model. The main objective of this model is to apply secured privacy preserving model in supervised decision tree classifier to protect the decision patterns in large datasets. In this model, a novel cubic decision tree measure is optimized to construct the optimal decision tree patterns on the given dataset. Proposed model is usually designed to improve the privacy in private data in data mining and to predict accurate decision rules on high dimensional data. In the proposed model, data from different sources are integrated with policy identifiers as key to construct decision patterns using the proposed attribute selection measure. Here, each data source key is mapped with the decision tree model to form the decision patterns. These decision patterns along with key identifiers are used to hide the privacy patterns among the integrated data sources.

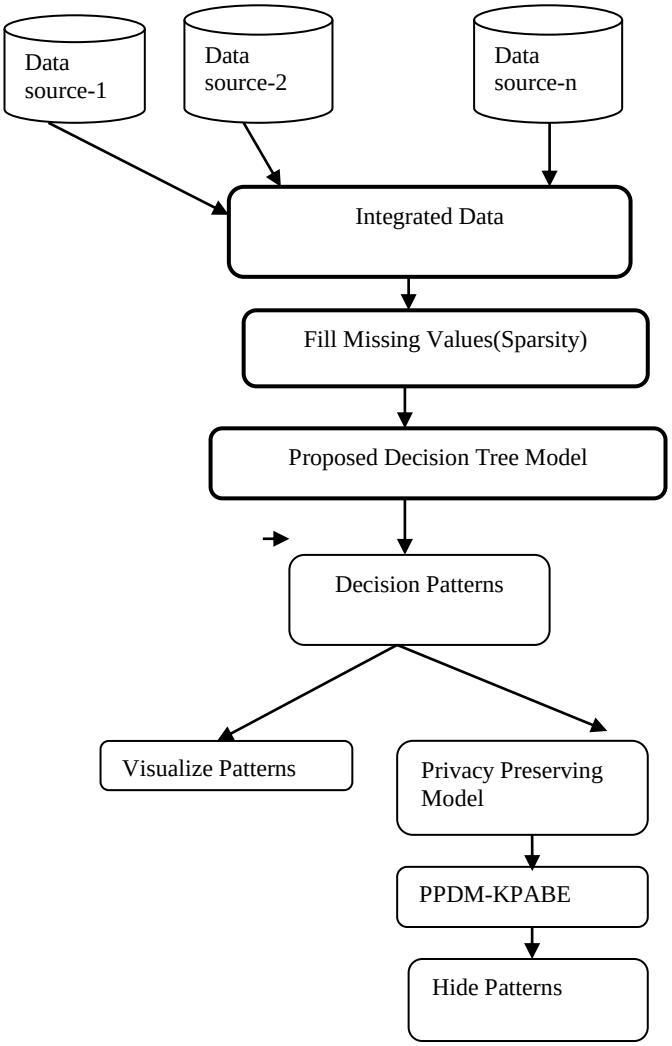


Figure 1: Proposed Privacy Preserving Framework

Figure 1, describes the use of privacy preserving model for high dimensional private datasets. Initially, centralized database is created from different sources for privacy preserving. Proposed model considers high dimensional distributed data as input for data pre-processing. Proposed model was implemented in two main phases, first phase is data filtering and the second phase is privacy preserving model using decision tree technique.

Filter Based Privacy Preserving Model:
Input :
Train Data: $D(x_1, x_2....x_n)$
Procedure:
For each attribute A{ $D(x_1, x_2....x_n)$ }
Do
if(isnumeric(A{ $D(x_1, x_2....x_n)$})
then
For each training instance $I(x_1, x_2....x_n)$
Do
If($I_A(i)$ ==null)
Then

$$V(i) := \frac{Max\{I_A\} - Min\{I_A\}}{\mu_A}$$

Done
if((A{ $D(x_1, x_2....x_n)$} is nominal)
Freq:=Frequency(D); //most occurred frequency value in the attribute
V(i)=V(F[i]) //Value of most frequent attribute value
Done

Proposed Privacy Attribute ranking measure:
Proposed attribute ranking measure is used to quantify the similarity between the positive class attribute to the negative type attribute using class distributions. It is basically derived from F-divergence. In our feature selection measure, we have used a novel cubic polynomial. The computation formula used to measure the attribute ranking measure is given in equation (3) as

$$\text{AttributeRanker} = \sqrt{(\sum_{i=1}^{m} \sqrt[3]{D_i / |D_i|} - \sqrt[3]{D_j / |D_j|})^2} \quad \text{-----(3)}$$

D_i is the positive class data

D_j is the negative class data

// Decision Tree construction
For each instance in D
If $D[i]$ ==Null
Then
Return leaf node with empty set.
Else if class($D[i]$)==m
Then
Return leaf node with class m.
Else
Split D into k disjoint partitions using stratified random sampling distribution where k=m(classes).
For each partition do
Partition the Dataset D into k uniform subsets as
$p(k) := D_r(i) / m$; // add instance to the kth partition
// user specified threshold for classification
$\lambda := userinput$;
// Get all ranked attributes greater than λ
Find the attribute ranking using the following equation
ρ ($A_i(p)$)= Apply proposed cubic attribute ranking measure
Where i,j=1,2….n attributes and m=number of classes.
End for
For each ranked instance in ρ

Do

If($\rho(A_i(p)) > \lambda$)

Then

Apply PPDM-KPABE($\rho(A_i(p))$);

End if

Done

Select the root node in the tree pattern using the attribute with highest ARank in all the partitions.

This process is repeated until no more instances in the partitions.

Privacy Preserving Decision Patterns using KP-ABE(KeyPolicy Attribute Based Encryption):

Every individual user is associated with an access tree structure. Threshold gates represent the nodes of access trees, whereas attribute values are assigned to leaf nodes. User's secret key has the responsibility of representing access tree structure. Cipher text includes set of attributes and the secret keys are merged with monotonic access structures in order to choose the particular cipher text for that partition to decrypt. PPDM KP-ABE is implemented in the applications of one-to-many communications. It includes four distinct phases which are mentioned below

Setup : It accepts input K as a security parameter and generates PK as public key. The procedure also generates a master secret key MK. The process of encryption requires public key PK. MK is responsible for the production of user secret keys and the authority has rights to access these secret keys.

Encryption : The encryption algorithm accepts plain text original message M, generated public key PK along with a set of attributes and policies as input. The algorithm produces ciphertext E which is the encoded version of the original message M.

Key Generation: Key generation algorithm includes an access structure Ta long with master secret key MK as input. The objective of key generation algorithm is to generate a secretkey SK. SK is responsible for decryption of the encrypted cipher text using the set of attributes matches T.

Decryption: The decryption algorithm accepts user's secret key SK satisfying the access structure T and cipher text E as input. The algorithm results the original plain text message M, when the attribute set satisfies T.

Key Policy-Attribute Based Encryption method can be applied to get fine-grained access control along with extended flexibility in order to have control on users. Hence, this scheme is better than that of classical attribute based encryption scheme.

4. Experimental Results

Experimental results are implemented on 2.4GHz Intel core i5 processor with minimum 2GB RAM. We used Jama, JUnit, ABE third party libraries for the proposed privacy preserving model. We used KDD99 full dataset and 10 fold cross validation for the testing purpose. In 10-fold cross validation, entire training data are randomly divided into 10 equal sub- partitions. This test data is used to estimate the proposed model accuracy.

Dataset Description

Table 1 describe the diabetes and KDD dataset used as well as their attributes, size and number of classification class labels. In 1999, the KDD acknowledged and approved DARPA data as the conventional benchmark data base for IDS called KDDCup99 which is available in http://www.kdd.ics.uci.edu/databases/kddcup99. In KDD99 data, each instance is associated with different attribute values with labelled attack names. These KDD99 class labels are classified into 5 types namely-normal, Probe attack, DOS attack, U2R attack and R2L attack. A complete KDD99 dataset contains nearly 4 million labelled instances with 41 features that are divided into 22 different types of attacks and is summarized in Table 1.

Table 1 : Details of labelled attacks

Attack Category	Attack Name
Denial of Service	Smurf, Neptune, back, land, pod,teardrop
Probe	IPsweep, satan, portsweep, satan
R2L	Ftpwrite,Imap,Guesspassword,warezclient
U2R	Perl, Bufferoverlow, Rootkit

Table 2: Performance analysis of Proposed model to the existing models on KDD Dataset

Model	Patterns	Execution Time
NaiveBayes+ABE	533	8355
SVM+ABE	325	9384
Logistic+ABE	353	8352
DecisionTree+ABE	643	7843
Homomorphic+PPDM	735	4953
Improved decision tree+Homomorphic	713	4724
IPPDM+KPABE	683	4012

Table 2 describes the performance analysis of proposed model to the existing privacy preserving model on KDD dataset. From this table, we can observed that proposed model has less computation time compared to traditional models.

Table 3: Performance of the proposed model to the existing model using homomorphic PPDM

Model	Patterns	Privacy Time	RecoverTime
NaiveBayes +ABE	533	8355	6583
SVM +ABE	325	9384	8647
Logistic +ABE	353	8352	7574
DecisionTree +ABE	643	7843	6966
Homomorphic +PPDM	735	4953	5147
Improved Decisiontree +Homomorphic	743	4525	4685
Proposed Model(IPPDM+KPAB)	683	4012	4214

Table 2 describes the performance analysis of proposed model to the existing privacy preserving model on KDD dataset. From this table, we can observed that proposed model has less privacy and recover time compared to traditional models.

Model	Patterns	Classification Accuracy
NaiveBayes+ABE	533	0.864
SVM+ABE	325	0.913
Logistic+ABE	353	0.925
DecisionTree+ABE	643	0.9413
Homomorphic+PPDM	735	0.8945
Improved Decision-tree+Homomorphic	743	0.9734
Proposed Model(IPPDM+KPAB)	683	0.9825

5. Conclusion

As the size of decision patterns over the high dimensional datasets is increasing the time taken to preserve the patterns using traditional models also increases due to high dimensionality and size of the data. Therefore, a novel privacy preserving model on high dimensional data is required to handle large data. In this paper, a novel privacy preserving based classification model was designed and implemented on large datasets. In this model, a filter based privacy preserving model using improved decision tree classifier is implemented to preserve the decision patterns using IPPDM-KPABE model. Experimental results proved that the proposed

model has high computational efficiency compared to the traditional privacy preserving model on high dimensional datasets.

References

[1] P. K. Fong and J. H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets," in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 2, pp. 353-364, Feb. 2012.

[2] L. Liu M. Kantarcioglu B. Thuraisingham "Privacy Preserving Decision Tree Mining from Perturbed Data" Proc. 42nd Hawaii Int'l Conf. System Sciences (HICSS '09) 2009.

[3] M. Shaneck Y. Kim "Efficient Cryptographic Primitives for Private Data Mining" Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS) pp. 1-9 2010.

[4] L. Xu, C. Jiang, Y. Qian, J. Li, Y. Zhao and Y. Ren, "Privacy-Accuracy Trade-off in Differentially-Private Distributed Classification: A Game Theoretical Approach," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1.

[5] P.K. Fong Privacy Preservation for Training Data Sets in Database: Application to Decision Tree Learning 2008.

[6] J. Vaidya C. Clifton "Privacy Preserving Association Rule Mining in Vertically Partitioned Data" Proc Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02) pp. 23-26 2002-July.

[7] L. Xu, C. Jiang, Y. Qian, J. Li, Y. Zhao and Y. Ren, "Privacy-Accuracy Trade-off in Differentially-Private Distributed Classification: A Game Theoretical Approach," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1.

[8] M. Bewong, J. Liu, L. Liu, J. Li and K. K. R. Choo, "A Relative Privacy Model for Effective Privacy Preservation in Transactional Data," 2017 IEEE Trustcom/BigDataSE/ICESS, Sydney, NSW, 2017, pp. 394-401.

[9] A. Gkoulalas-Divanis G. Loukides J. Sun "Publishing data from electronic health records while preserving privacy: A survey of algorithms" JBI vol. 50 pp. 4-19 2014.

[10] H. Zakerzadeh C.C Aggarwal K. Barker "Managing dimensionality in data privacy anonymization" KIS vol. 49 no. 1 pp. 341-373 2016.

[11] S. Qiu, B. Wang, M. Li, J. Liu and Y. Shi, "Toward Practical Privacy-Preserving Frequent Itemset Mining on Encrypted Cloud Data," in IEEE Transactions on Cloud Computing, vol. PP, no. 99, pp. 1-1.

[12] Sastry, J.K.R., Ganesh, J.V., Bhanu, J.S., I2C based networking for implementing heterogeneous microcontroller based distributed embedded systems, Indian Journal of Science and Technology, Volume 8, Issue 15, 2015

[13] Sastry, J.K.R., Naga Sai Tejasvi, T., Aparna, J., Dynamic scheduling of message flow within a distributed embedded system connected through a RS485 network, ARPN Journal of Engineering and Applied Sciences, Volume 12, Issue 9, 1 May 2017, Pages 2809-2817

[14] Sastry, J.K.R., Suresh, A., Bhanu, S.J., Building heterogeneous distributed embedded systems through rs485 communication protocol, ARPN Journal of Engineering and Applied Sciences, 2015, 10(16), pp. 6793-6803