International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 10, Number 1 (2017)

An Efficient Framework for Privacy Preserving in Correlated Data based on Tree Distribution Approach

Radhika A. Gore

M.E., Department of Computer Science & Engg,

D. Y. Patil College of Engineering & Technology, Shivaji University, Kolhapur,

Maharashtra, India.

G. A. Patil

H.O.D, Department of Computer Science & Engg,

D. Y. Patil College of Engineering & Technology, Shivaji University, Kolhapur,

Maharashtra, India.

E-mail : [1] patilradhika38@gmail.com [2]gasunikita@yahoo.com

Abstract

Privacy preserving in data mining plays a vital role in assuring that one or more user enters into the system with high privacy. Many times attributes in a dataset are correlated with each other, therefore chances of data leakage are more. So issues regarding privacy have occurred. In this paper, privacy has been provided to correlated dataset in which user fires a query to get the output. In general, attributes in a dataset are sampled independently. However, In real-world, attributes in a dataset are rarely independent. If one attribute is dependent on another, then it leads to privacy violation. The solution to this problem is to provide privacy using correlated tree distribution approach in which query gets scrutinized to get proper attributes from the dataset. Different steps are applied like K-Means Clustering algorithm, Correlation Computation and Tree Creation. If-Then Rules are used to set ranges in order to display only the data needed by the user, hiding other sensitive information from the user. Access control policies are provided in order to protect information. Many existing systems have a drawback in which the major drawback is that all have huge time complexity. In tree distribution approach, the drawback has been overcome by strictly focusing on finding correlation amongst data. To enhance the privacy in Non-IID (Independent and Identically Distributed) datasets, proposed system uses tree based traversing technique for maintaining a strict access control over the attributes. The main aim of the system is to provide privacy when user searches for the data globally.

Keywords: Ontology, Person correlation, K-Means clustering, Correlation tree, Cosine similarity, Fuzzy Logic.

Introduction

In general, records in a dataset are worked independently, but if considered as per the depth of a particular dataset, it seems that data in the dataset are dependent on each other. In this case, particular user fires a query on a dataset expecting the output as per query, but if data is dependent on any sensitive attribute, then it leads to leakage of the particular information. It is necessary to find out correlated attributes so that privacy can be applied to particular attributes. There are different kinds of queries user can design to get data, so careful examination of query in detail is necessary. In many existing system, access control policies are not provided which again leads to privacy issue.

Knowledge discovery [1] in databases refers to extracting knowledge from data in the context of large databases. For this, various data mining tasks such as artificial intelligence, machine learning, statistics and many other techniques are used. Data mining is used to find patterns and relationships in the data. Privacy preserving in Data Mining is the study of achieving some data mining goals without scarifying the privacy of individuals. Privacy preserving in distributed data has various applications. Each application has different

conditions: What is meant by privacy? What are the desired results? how is the data correlated and distributed with each other? What is the highest distribution factor in dataset? etc. Data mining can extract knowledge from large data collections, sometimes these collections are divided among various parties. Privacy may prevent the parties from directly sharing the data and some types of information about the data. This paper presents some mechanism and shows how they can be used to solve several privacy-preserving data mining problems.

Correlated Iteration Mechanism is used for finding correlation between attributes. However, this mechanism takes more time to compute correlation factor. It also considers non-correlated attributes which are not necessary because privacy has been given to only correlated data and it also leads to space complexity problem. Access control policies are also poor. In some system there may also exist a mismatch between the estimated prior statistics and the true prior statistics, due to a small number of observable samples. The proposed system aims to find strict correlation amongst attributes in the datasets in which important data can be protected.

The remaining section of this paper is organized as follows: Section II includes related work. Section III includes Proposed Methodology. Section IV includes Results and discussions and Section V includes conclusion and future scope.

Related work

Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao [2] have proposed a privacy-preserving outsourced frequent itemset mining solution for vertically partitioned databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, they built a privacy-preserving outsourced association rule mining solution for vertically partitioned databases. They provide solutions to protect data owner's raw data from other data owners and the cloud. Also ensure the privacy of the mining results from the cloud.

Tianqing Zhu, Ping Xiong, Gang Li [3] have proposed Correlated Iteration mechanism to answer large number of queries, which proposed a correlated data privacy for hiding information in NON-IID dataset. It successfully prevents the utility and enhances privacy guarantee while preserving privacy. It saves the privacy budget and decreases the noise for each query. Records in the datasets are often correlated with each other, and this may reveal extra information.

YilinShen and Hongxia Jin [4] have proposed relaxed admissible mechanism to achieve recommendation accuracy and privacy preservation. With the help of this mechanism two contradictory goals they met i.e. recommendation accuracy and privacy preservation. User's private data is required which put users at risk. User's perturbed data can be used in existing recommender system in order maintain privacy. Relaxed admissible mechanism is used to achieve feasible and useful perturbations. It benefits companies with high quality personalized services and strong privacy protection on perturbed data.

Anil Pratap Singh and AbhishekMathur [5] have proposed Chaotic system based data perturbation technique with multilevel trust. It is relatively very difficult to crack the privacy of the original data offered by the proposed method when compared with random perturbation based techniques especially when the number of published perturbed copies of same trust level increases. Chaotic system generates highly unpredictable noise; it contains a couple of differential equations and need to be solved for each point of the required number of solution points. The solution of these equations may be time consuming.

Pui K. Fong and Jens H. Weber-Jahnke, [6] have proposed a approach called new privacy preserving approach via dataset complementation. It provides private information from the samples while keeping their utility. This approach converts the sample data sets into some unreal data sets so that any original data set is not reconstructed if a theft were to steal some of the contents. If training data sets are revealed, privacy preservation via data set complementation will not work. To overcome this limitation, a cryptographic privacy in preserving approach along with data set complementation can be used.

Haibat Jadhav, Prof. Pankaj Chandre [7] have proposed one of the significant and popular data mining process is association rule mining which is used to find frequent patterns in the given dataset in which the apriori algorithm are the most common for mining frequent item set. This paper explored DES algorithm at client side for generating the secret key to encrypt the items of the support table. Apriori algorithm is used at server side for getting transaction data from the client side by applying threshold or sigma value to filter out the item set whose frequency is less than sigma value. The main focus of this paper is to achieve more security of the client side.

Julius Adebayo, Lalana Kagal [8] have proposed a transformation procedure for large scale individual level data that produces output data in which no linear combinations of the resulting attributes can yield the original sensitive attributes from the transformed data.

Iyer Chandrasekharan, P.K. Baruah Prashanti Nilayam [9] have proposed a novel hybrid method to achieve k-support anonymity based on statistical observations on the datasets.

Proposed Method

The overall technique works as follows: First, user enters their requirement for the data in relational database as a query. Next data mining tasks are performed such as stemming and stop word removal. Using Noun detection algorithm, noun words and top words are identified. In ontology, hierarchical structure of parent and child keyword is created. Next, ontology reads complete dataset and clusters are formed using K-means clustering algorithm. After formation of clusters, Pearson coefficient is calculated using Pearson correlation algorithm. Based on these Pearson values tree is created using preorder tree traversal. Using cosine similarity, node clusters are formed. In the fuzzy logic, If- then rule is applied for all values and classification is done. In the Key management, privacy is applied and information is displayed based on the user requirement.

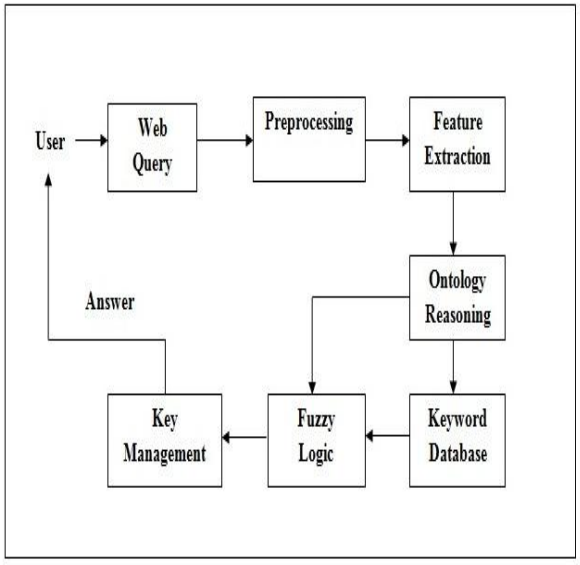

Figure 1: Overview of the proposed work

Figure1 shows the framework for access control mechanism over the relational database.

Let 'S' be the system for privacy preserving for NON-IID dataset which contains input(I) as query from user. Output (O) will be the result data with privacy. Let T_n be the data set, Q_n be the set of query keywords on the basis of users query. The correlation factor is being calculated in the datasets in which numbers of records are considered as $x_1, x_2 \ldots x_n$ which belongs to T_n. Privacy can be preserved on the basis of the kind of query user is firing. So, Q_n is considered set of query keywords which are related to sensitive information present in the dataset. Different processes are used to find the strict correlation factor.

1) Preprocessing

User enters their requirement for the data in relational database as a query and then query is passed to the preprocessing. In the preprocessing, stop words are removed and stemming is done. Preprocessing is done on query given as input in which special symbols (.,;') are removed. Stop words (is, are, was, were, it, but) are removed. Stemming is used to get proper word without any 'ing', 'ly', 'ed'. Output is a set of preprocessed data attributes.

2) Feature Extraction

In the Feature Extraction technique new feature is obtained from the original features using various statistical techniques. This system makes use of two different features, which is extracted from query entered by the user.

Noun detection

Noun extraction from the entered query plays a vital role in identification of the perception of the query. This is done by comparing each word of the query with the dictionary collected for almost 1, 00,000 words of English language.

Term Weight Identification

The most repetitive words in text are most important words. So the system will identify the list of most repeated words and will consider some top n elements (where n is user defined) as the important words for text to store in vector.

Input to this step is preprocessed query.
Output is an extracted feature.
Features get extracted in which occurrence of attributes get calculated. Also pronoun from query gets extracted.

3) Ontology Reasoning

In the Ontology Reasoning, yielded independent queries are very much linked with the mother queries. Ontology first reads hierarchy, then it decides relations, then it identifies master keywords and finally stores those words in keyword vector.

Attribute Identification

Dataset attributes are identified for the query entered by the user using first part of the ontology reasoning.

This process is done by matching the dataset attribute with attribute names.

Cluster computation

Identified attributes are set for clustering using K-means clustering process. K-means clustering is one of the unsupervised computational methods used to group similar objects in to smaller partitions called clusters.

Correlation Identification

In this step, clusters are formed based on the attributes obtained from the attribute identification step using K-means algorithm. These clusters are used to calculate correlation between the attributes by the means of Pearson correlation technique. This can be shown as follows.

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

N = number of pairs of data
$\sum xy$ = sum of the product of paired data
$\sum x$ = sum of x data
$\sum y$ = sum of y data
$\sum x^2$ = sum of squared x data
$\sum y^2$ = sum of squared y data

Algorithm : Pearson Correlation
//Input: Two parameter matrix of N rows and 2 columns and Let matrix be M
//Output: Pearson factor (i.e. in between 0 to 1)
1: Calculate sum of square of column 1 as SS1
2: Calculate sum of square of column 2 as SS2
3. Calculate square of mean of column 1 as m1
4. Calculate square of mean of column 2 as m2
5. Calculate square root of SS1-m1 as SQ1
6. Calculate square root of SS2-m2 as SQ2
7. Calculate denominator as DR as SQ1 * SQ2
8. Calculate sum of column 1 as sum1
9. Calculate sum of column 2 as sum2
10. Calculate product of sum1 and sum2 as TP
11. Calculate Mean product as MP as TP/ N
12. Calculate sum of product of all rows as PS
13. Calculate nominator as NR as MP*PS
14. Calcualte pearson coefficient as NR/DR
15. return pearson coefficient

Correlation tree

Clusters and Pearson coefficients are presented in table format. Tree is created using preorder tree traversal based on the following algorithm

Algorithm **:** Tree Traversal (Cluster Vector F_{vi})

Begin:
Create an empty tree as T
Create the Root Node for first cosine similarity R_n
For each element of clusters F_v
Compare the distance with the root node
If (F_{vi}Cosine Similarity <R_n)
Add node as left child in T
Else
Add node as Right child in T
End For
End

Cosine Similarity

Cosine similarity metric is frequently used when trying to determine similarity between two entities. In this similarity metric, the attributes (or words, in the case of the entities) are used as a vector to find the normalized dot product of the two entities. By determining the cosine similarity, node clusters are formed. Cosine similarity is expressed using the following mathematical equation:

$$similarity(x, y) = cos(\theta) = \frac{x \cdot y}{||x|| * ||y||}$$

4) Fuzzy Logic

The Fuzzy Logic involves the following steps:

Fuzzifier

In the fuzzifier, crisp inputs are taken, which are result of the performance parameter and after fuzzification, the inference engine was referred to the rule base containing fuzzy IF-THEN rules.

If-Then Rule

In this step, we consider the final performance score obtained from fuzzifier. In inference engine, the most important part is the definition of fuzzy IF-THEN rules. The essential parameters are extracted from these rules according to database performance criteria.

Fuzzy logic on this are used to classify and extract the cluster data having correlation in the following steps

a) First collect all leaf nodes of the tree in a set.
Then S= { { h,0.2}, {i,0.26}, {j,0.29} , {k,0.32}, {l,0.4}, {m,0.43}, {n,0.46}, {k,0.6}}
b) Sort all the subset elements in descending order where higher cosine similarity elements were in the higher priority.
c) System records the lower and upper similarity elements, like low=0.2 and high=0.6

d) Fuzzy crisp values are created based on the following criteria as shown below:

VERY LOW - 0.2 to 0.28
LOW - 0.29 to 0.36
MEDIUM - 0.37 to 0.44
HIGH - 0.45 to 0.52
VERY HIGH - 0.53 to 0.6

e) By traversing all nodes, every node's fuzzy range is recorded like {h,0.2,VERY LOW}, {a,0.4, MEDIUM} in the fuzzification process.

5) Key Management

Query for each keyword stored in the keyword vector is executed. Then results of each keyword are merged to get final result. The final answer is obtained by filtering the merged result and each user is assigned different access control based on their profiles. By applying IF-then rules for all the fuzzy ranges and then privacy was applied according to that on the data.

Results and Discussion

Experimental evaluation is carried on system of privacy preservation on non IID dataset which is developed with the approach of tree pattern analysis. Proposed system is developed on java based windows machines which uses Netbeans as IDE. System is put under hammer for crucial testing for its authenticity as mentioned in below tests.

For testing the performance of proposed system; Precision, Recall, F-measure are computed.

Table 1: Precision, Recall, F-Measure of Correlated tree distribution approach and Association rule mining approach using Adult dataset for three attributes.

	Precision	**Recall**	**F-Measure**
CTDA(Proposed)	60%	100%	75%
ARMA(Existing)	33%	100%	49.62%

Table 2: Precision, Recall, F-Measure of Correlated tree distribution approach and Association rule mining approach using Adult dataset for four attributes.

	Precision	**Recall**	**F-Measure**
CTDA(Proposed)	68.47%	100%	81.28%
ARMA(Existing)	47.24%	100%	64.16%

Table 3: Precision, Recall, F-Measure of Correlated tree distribution approach and Association rule mining approach using Adult dataset for five attributes.

	Precision	**Recall**	**F-Measure**
CTDA(Proposed)	85.56%	100%	92.21%
ARMA(Existing)	79.96%	100%	89.26%

The Privacy Preserving using Tree Distribution Approach system gives privacy to dataset as how sensitive information a user ask to system. The main aim of the PPTDA system is to find the correlated information. On providing access control mechanism user will get output. The Privacy Preserving using Tree Distribution Approach system is compared with Association Rule Mining Approach. In PPTDA system 'Adult' dataset is used for computation.

The ARMA system [2] implements privacy preserving using association rule mining and frequent item set mining gives the precision 33%, recall 100% and F-measure 49.62%.

The privacy preserving using tree distribution approach system gives better results than the association rule mining approach such as precision 42%, recall 100%, F-measure 59.15% as shown in Figure 2.

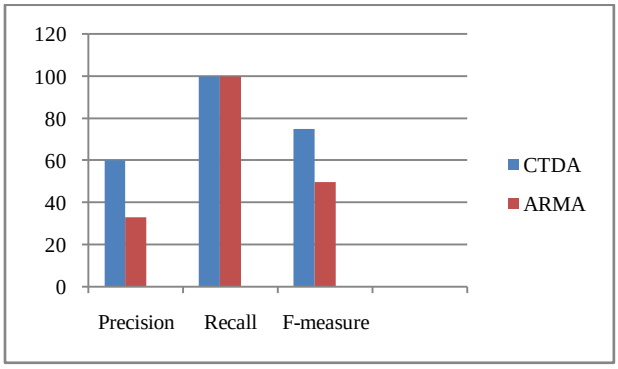

Figure 2: Graph showing Precision, Recall and F-measure for 3 Attributes.

Likewise different identified attributes are considered i.e. as shown in the figure 3. The privacy preserving for four attributes is shown.

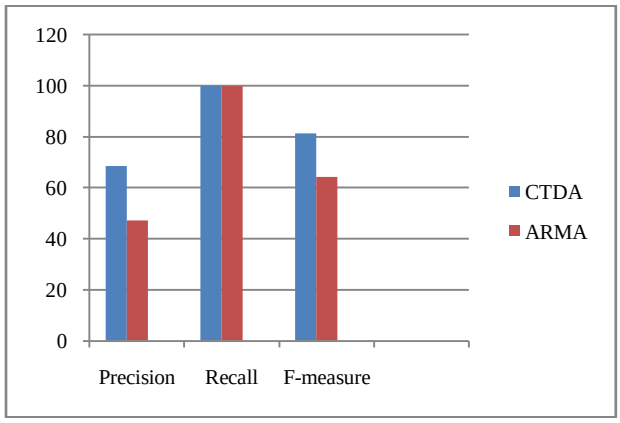

Figure 3: Graph showing Precision, Recall and F-measure for 4 Attributes.

The privacy preserving for five attributes is shown in figure 4.

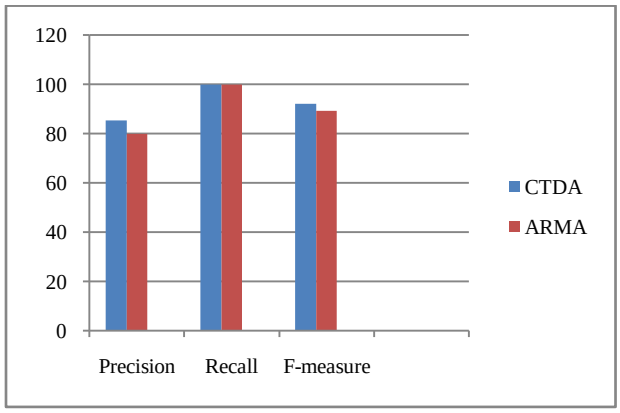

Figure 4: Graph showing Precision, Recall and F-measure for 5 Attributes.

Conclusion

This system gives comprehensive and general approach named correlated tree based approach for discovering informative knowledge in complex dataset for any kind of data mining applications. In proposed method, Access control policies are provided which is not provided in the existing system. In Privacy preserving tree distribution approach (PPTDA) system, privacy has been given to identify attributes based on the correlation amongst them, but in the ARMA system [2] the attributes are scrutinized using frequent itemset mining wherein only identified attributes gets privacy. The main aim of PPTDA (proposed system) is to find the Correlated information and provide privacy. Also different mechanisms as well as algorithms have been proposed which show how the sensitive information can be protected. The system is evaluated with the help of Precision, Recall and F-Measure.

In future, more work is needed concerning improvement in privacy. It can be applied for new applications. Although the techniques and algorithms used for preserving privacy are advancing fast, however, a lot of problems in this field of study remain unsolved. System can be enhanced to work on heterogeneous dataset.

References

[1] Zahid Pervaiz, Walid G. Aref and Nagabhushana Prabhu "Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data" IEEE Transaction On Knowledge And Data Engineering, vol. 26, no. 4, April 2014.

[2] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao , "Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases" IEEE transactions on Information Forensics And Security, Volume:11, May 2016.

[3] Tianqing Zhu, Ping Xiong, Gang Li, "Correlated Differential Privacy: Hiding Information in Non-IID Dataset" Information Forensics and Security, IEEE Transactions on, Volume:10, Feb. 2015.

[4] YilinShen and Hongxia Jin, "Privacy-Preserving Personalized Recommendation: An Instance-based Approach via Differential Privacy", IEEE International Conference on Data Mining, 2014.

[5] Anil Pratap Singh and AbhishekMathur , "A Chaotic Based approach for Privacy Preserving Data Mining Applications with Multilevel Trust", 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 978-1-4673-6126-2/13/$31.00 c 2013 IEEE.

[6] Pui K. Fong and Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree LearningUsing Unrealized Data Sets", IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 2, FEBRUARY 2012.

[7] Haibat Jadhav, Prof. Pankaj Chandre, "Association Rule Mining Methods for Applying Encryption Techniques in Transaction Dataset", 2016 International Conference on

Computer Communication and Informatics (ICCCI -2016), Jan. 07 – 09, 2016 .

[8] Julius Adebayo, Lalana Kagal, "A Privacy Protection Procedure for Large Scale Individual Level Data", 2015 IEEE.

[9] Iyer Chandrasekharan, P.K. Baruah Prashanti Nilayam," Privacy-Preserving Frequent Itemset Mining in Outsourced Transaction Databases" , International Conference on Advances in Computing, Communications and Informatics (ICACCl), 2015.

[10] Ali Makhdoumi , Nadia Fawaz,"Privacy-Utility Tradeoff under Statistical Uncertainty", Fifty-first Annual Allerton Conference 2013.

[11] RahenaAkhter, RownakJahanChowdhury, Keita Emura, Tamzida Islam, Mohammad ShahriarRahman, NusratRubaiyat, "Privacy-Preserving Two-Party k-Means Clustering in Malicious Model", IEEE 37th Annual Computer Software and Applications Conference Workshop, 2013.

[12] M Balamurugan, J Bhuvana and S ChenthurPandian, "Shared and Secured Data Partitioning For Privacy Preserving of Collaborative File Transfer in Multi Path Computational Mining ",978-1-4673-1989-8/12/$31.00 ©2012 IEEE.

[13] Xiaoyan Zhu, Momeng Liu, and Min Xie, "Privacy-Preserving Affinity Propagation Clustering over Vertically Partitioned Data", Fourth International Conference on Intelligent Networking and Collaborative Systems, 2012.

[14] C. Gokulnath, M.K. Priyan, E. Vishnu Balan, Prof. K.P. Rama Prabha and Prof. R. Jeyanthi, "Preservation of Privacy in Data Mining by using PCA Based Perturbation Technique" 2015 International Conference on Smart Technologies and Management 978 -1-4799-9855-5/15/$31.00 ©2015 IEEE.

[15] Jun Zhang,Yang Xiang, Yu Wang, WanleiZhou,Yong Xiang, and Yong Guan, "Network Traffic Classification Using Correlation Information", IEEE Transactions on parallel and Distributed on Parallel Systems, Vol. 24, No. 1, January 2013.