ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

ANALYSIS OF THE ALGORITHM AND TECHNIQUES FOR PRESERVING PRIVACYONDATA MINING

M.S.Saravanan*, V. Shanmukha Reddy
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Chennai, India.
IV CSE – B, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Chennai, India.
Email: saranenadu@gmail.com

Received on 10-08-2016 *Accepted on 06-09-2016*

Abstract

Now-a-days preserving privacy on data mining is one of the important problems in vast available personalized data. This paper defines preserving privacy ondata mining (PPDM) problem. It describes some of the new and rapidly emerging field of research in privacy preserving data mining and some exists problems. This paper gives an outlook of some classification methods for preserving the privacy on mining – data distortion method, data protection and some of the well-known PPDM algorithms –association rule mining, EM-clustering, ID3 for decision tree. Full evaluation of PPDM algorithm was illustrated; these algorithms are usually a modification of a well-known data mining techniques with some preserving privacy techniques including data utility, algorithm performance, degree of privacy protection, difficulty of different data mining. This paper will give a summary and high level overview of the privacy preserving data mining (PPDM).

Keywords: Preserving privacy, PPDM, ID3 decision tree, clustering data, reconstruction, block mining.

1. Introduction

Data mining refers to the means for mining knowledge (data) from huge volumes of data. It is also popularly termed as KDD (knowledge discovery from data). At first data mining works on data ware house model in which all the data are gathered together at one central place. In this model only one party is owned the entire data, who will be responsible for use and generate the data without disclosing to the third party[1]. Moreover invarious real time applications of this technology, privacy concerns might prevent this. The top issue may be that to give the personal ID's the data may contain the list of attributes which should be efficient. Vertical and horizontal data's can split across multiple parties where none of the party is permissible to transfer data from one another site and having the

same restrictions different parties are having the different attributes of the data. And Hence, restrictions and some rules are restrictedin the use of data mining models. PPDM has emerged to solve this issue by using various protection algorithms. Privacy protection method is decided by the application of different privacy protection, which is use to preserve the data at lower privacy level by using probabilistic and statistical models. The main purpose of privacy protection is to preserve the privacy of a data. Data release based privacy protection is used to provide common protection in many applications and it I used to make the privacy protection algorithm more versatile.

Privacy protection research is focused on various areas including data distortion, data utility, data encryption, vertically and horizontally partitioned data, EM clustering, data release, association rules mining, decision tree mining, naïve Bayes. This paper reviews privacy protection arising issues and privacy protection algorithms.

2. MAIN METHODS AND ALGORITHMS USED IN THE PRIVACY PRESERVING DATA MINING

There are many methods of data mining privacy protection; our classification of privacy preserving data mining is based on four aspects which are briefly described as:-

Distributed data:

The data are vertically and horizontally partitioned according to that privacy protection algorithm is execute. Different records in different sites are may be vertically or horizontally data each database record attribute values in different sites.

Distorted data: In this method, it first modifies the data before release so as to provide the privacy protection and it includes blocking, merging, sampling, swapping, perturbation. All this can be achieved by modifying the attribute value.

Algorithms for data mining: the classification of data mining include decision tree, clustering, Naïve Bayes, association mining rule etc.[2]

Hidden data or rules: This method is used to hide the data or rules of original data because the rules of original data are very complex.

Privacy protection: In order to provide protection to the privacy of data, there need to carefully modify the data to achieve the high data utility. Reasons for doing this are:-

(a) Minimizing the information loss in data by modifying the selected values.

(b) Providing the secure multiparty computation by encryption technologies.

(c) Data reconstruction can be done at the last so as to reconstruct the original data distributed from random data.

3. Techniques for Privacy Protection

3.1. Privacy preserving distributed mining

In the environment of privacy preserving mining, people use the encryption based approach to preserve their data. On the basis of cooperation two or more parties mine their data but none the party is willing to reveal their data. This is a multiparty secure computation, which focuses on to convert the methods of data mining into multiparty secure computation issues, such as data generalization,data clustering, data aggregation, association mining rules. Multiparty secure computation includes the secure set union, secure set sum, secure size of set intersection, scalar product. Let us discuss association mining rules.

Association mining rules for horizontally partitioned data: The transactions are distributed in *n* sites in a horizontally distributed database. The sum of all local counts will be equalto the total support count of itemset.

Association mining rules for vertically partitioned data: In a vertically partitioned database the data set different attributes for each itemset in different sites. If the itemset's support count is calculated in secure manner, then we can check whether the support is greater than threshold, and check whether it is frequent.

3.2. Techniques for data distortion

In order to provide the protection to the released data, many people use lots of data mining technology to hide their data. The main purpose of privacy protection is (a) Hide the sensitive data from the original data, (b) original data and hide data have the same characteristic, (c) have the same accuracy as the original data. The algorithms for data mining such as clustering, association mining rules, classification, need to choose data to modify and the choice of modified data is a hard NP problem. To deal with this complex problem distortion technique is use such as condensation, blocking and random perturbation.

Mining association rules based on perturbation:

Rules emergence in data set is judge by the statistical significance, and support and confidence as metric. User defined support and confidence are greater than or equal to all association rules, to pure the original data set association rules hiding technique is the following method.

(a) When the data is purified confidence and support is not allowed to appear and all sensitive rules can only appear on the original data mining.

(b) Original data can dug out in the clean data set at the same support and confidence and non-sensitive rules can dug out in original data set.

(c) Sensitive data cannot be dug out in the original data and original data cannot be dug out in the purification at the same support and confidence.

Using block mining association rules: Data block is another perturbation for association rules. In this block method, the data item's property value replace with the question mark, that using false value instead of actual value rather than using unknown value instead of actual value is very popular in medicine. Reference [3] proposed a method using blocking in association mining rules, which changes the definition of minimum support appropriately, replace with minimum support and minimum time interval. As long as the confidence of sensitive rules below the middle of the confidence interval, support of sensitive rule below the middle of support interval are there then the privacy is not violated.

3.3 Data Reconstruction Technology

Use of data perturbation or reconstruction in data convergence layer provides much PPDM technology. In Reference [4], use the individual records value perturbation as training data to construct a decision tree classifier. Since the original values of individual records cannot be accurately estimated, the author considers estimating accurately the original distribution. Bayesian method is considered in order to reconstruct the original distribution.

4. Algorithms for Privacy Preserving

Privacy preserving data mining (PPDM) has various algorithms, and our classification of PPDM algorithm based on four aspects which are briefly described as follows:-

4.1. Decision tree Mining

In the paper [5], the version ID3 decision tree for privacy preserving is described where DB1 and DB2 are the two parties with databases. On joint database DB1UDB2 applying decision tree algorithm without revealing any unwanted information about their database. In traditional ID3 algorithm, the data are partitioned by choosing the best attribute at each level of the tree. The tree will be completed when there is no data to split up and each data is partitioned individually into a single simple class value. The information gain theory is used in the selection of the best attribute and selects the attribute which maximizes the information gain and minimize the entropy of the partitions.

4.2. Association Rule Mining

This paper also describes, across multiple sites the privacy preserving technique for horizontally partitioned data sets. Let I=$\{i_1, i_2, ...i_n\}$ be a items of set and T=$\{T_1, T_2, ...T_n\}$ be a set of transactions where each $T_i \subseteq I$. A transaction T_i

contains an item set $X \subseteq I$ only if $X \subseteq T_i$. An association rule implication is of the form $X \Rightarrow Y$ ($X \cap Y = 0$) with support s and confidence c if %s of transaction in T contains $X \cup Y$ and c% of transactions that contains X and Y also. Transactions are distributed in *n*sites. The global support count of an item set will equals the sum of all local support counts. The global confidence in terms of global supportis[6]:

$$SUP_g(X) = \sum_{i=1}^{n} SUP_i(X)$$

$$CONF_g(X \Rightarrow Y) = SUP_g(X \cup Y) \Big/ SUP_g(X)$$

The main aim of the association mining rule in privacy preserving is to find all those rules with the global confidence and global support higher than the specified minimum confidence and support by the user.

4.3. EM Clustering

Based on the values of attributes the data are grouped together called "clusters" using the clustering technique.[7] EM algorithm is one of the well-known algorithms for clustering which works well with the discrete as well as continuous attributes. A version of privacy preserving algorithm in multi-site case with the data partitioned horizontally is described below.

Assuming the data is single dimensional (single attribute y) and are partitioned across sites. Each site has n data items ($n = \sum_{l=1}^{s} n_l$). Let $z_{ij}^{(t)}$ denote the cluster membership for the ith cluster and jth data point at the (t)th EM round. In the E step, the values μ_i (mean of i), σ_i^2 (variance of i) and π_i (estimated proportions of i) are calculatedusing the sum:

$$\sum_{j=1}^{n} z_{ij}^{(t)} y_j = \sum_{l=1}^{s} \sum_{j=1}^{n_l} z_{ijl}^{(t)} y_j$$

$$\sum_{j=1}^{n} z_{ij}^{(t)} = \sum_{l=1}^{s} \sum_{j=1}^{n_l} z_{ijl}^{(t)}$$

$$\sum_{j=1}^{n} z_{ij}^{(t)} (y_j - \mu_i^{(t+1)})^2 = \sum_{l=1}^{s} \sum_{j=1}^{n_l} z_{ijl}^{(t)} (y_j - \mu_i^{(t+1)})^2$$

The summation of the second part in all these cases is local to every site. It is clearly seen that sharing value does not reveal y_i to the other places. It is also not necessary to share the inner summation values and n, but just computing n and the global summation for the value above using the secure sum technique.

5. Conclusion

In this paper, it is focusing on the technologies used in the data mining with respect to preserving privacy.First we introduce the study of protection privacy status and main research methods and algorithms. By the above techniques and algorithms it is clear that without compromising with the accuracy of the original data it is potential to ensure privacy, and has a bound on the communication cost and computation.

References

1. Patrick vogel, torstengreiser, dirk Christian mattfeld“ understanding bike shring system using data mining: exploring activity patterns”, science direct, procedia social and behavioural sciences 20(2011) 514-523.
2. M.S.Saravanan, A.Rajalakshmi, ”Privacy Preserving Open Audit for Protecting Files in the Cloud Storage”, Published in International Journal of Applied Engineering Research by Research India Publications, India, Vol.10, Issue.33, May’ 2015, pp.24618-24621, ISSN:0973-4562.
3. L.chang and I. Moskowitz, “ An integrated framework for database privacy protection”, data and application security, springer boston, 2002, pp.161-172.
4. A. weichselbraun, s.gindl, A. scharl “ Enriching semantic knowledge bases for opinion mining in big data applications”, science direct,2014.
5. Caiyandai, Lin chen “An algorithm for mining frequent closed itemsets in data stream”, science direct, physics procedi 24(2012) 1722-1728.
6. J.M. Adamo,” Data mining for associational rules and sequential patterns: sequential and parallel algorithms”, springer-verlag,2001.