Privacy Preservation Of Data Sets In Data Mining

[1]Alvina Anna John [2]S.Deepajothi
[1]PG Scholar [2]Assistant Professor,Chettinad College Of Engineering & Technology,Karur

Abstract

Data mining is the process of extracting or mining knowledge from large databases. Privacy-preserving data mining plays an important role in the areas of data mining and security. In the area of privacy preserving data mining, the data mining algorithms are analyzed for their impact on data privacy. The goal of privacy preserving data mining is to develop algorithms to modify the original data set so that the privacy of confidential information remains preserved and as such, no confidential information could be revealed as a result of applying data mining tasks. The existing privacy preservation technique is the data set complementation approach and it fails if all data sets are leaked as the data set reconstruction algorithm is generic. The proposed method provides privacy preservation by converting the original sample data sets in to a group of unreal data sets and then applying cryptographic privacy protection to sensitive values. The cryptographic technique implemented is RSA. This method provides privacy preservation with improvement in accuracy. This work covers the application of new privacy preserving approach with the Naïve Bayesian classification algorithm.

1. Introduction

Data mining is a method that helps to extract useful information from large databases. It is the technique of extracting relevant data from giant databases through the utilization of data mining algorithms. As the quantity of information doubles each year, data mining is becoming an increasingly important tool to transform this data into information. Data mining deals with massive databases which can contain sensitive information. It needs data preparation which can discover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has enlarged the risks of revealing sensitive data. Providing security to sensitive information against unauthorized access has been a long term objective for the database security research community and for the government statistical agencies. Hence, the protection issue has become an important area of research in data mining. The releasing of personal data in its most specific state poses a threat to the privacy of an individual. The threat to an individual's privacy happens when anyone who has access to the newly-compiled data set is able to identify specific individuals. The disclosure of extracted patterns open up the danger of privacy breaches that may reveal sensitive information to malicious users thus causing privacy violation in data mining. Hence there is a need for privacy preservation in data mining. Privacy preservation in data mining (PPDM) is the process of providing security to sensitive data in a database against unauthorized access. PPDM is implemented in order to protect the sensitive information and to prevent violation of privacy.

2. Related Work

M. Kantarcioglu, et.al gives a discussion about the concept of privacy violation in data mining [10].This paper defines that privacy-preserving data mining has concentrated on obtaining valid results when the input data is private. Alexandre Evfimievski et.al refers Privacy-preserving data mining (PPDM) as the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure[2]. The term privacy-preserving data mining was introduced in the papers Agrawal et.al [3] and Lindell et.al[4]. These papers considered two fundamental problems of PPDM: privacy-preserving data collection and mining a data set partitioned across several private enterprises. Latanya Sweeney et.al proposed the k-anonymity as a model for protecting privacy[5]. This paper discusses k-anonymity protection and also examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. Ashwin Machanavajjhala et.al proposed the technique l-Diversity as Privacy preservation beyond k-Anonymity[6]. This paper provides a detailed analysis of the two attacks on k-anonymity called Homogeneity Attack and Background Knowledge Attack and provides a powerful privacy definition called l-diversity. Ninghui Li et.al gives a detailed study on a privacy preservation technique called t-Closeness as a Privacy Beyond k-Anonymity and l-Diversity[7]. This paper proposes a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. E. Poovammal et.al provides a detailed study on different privacy preservation techniques like data publishing, k-anonymity, l-diversity and a privacy preserving model[8].This paper discuss about the advantages and disadvantages of k-anonymity, advantage of l-diversity over k-anonymity and finally the privacy preserving model which performs privacy preservation based on the data type. If it is numerical data type, transformation is performed by categorical membership values and if categorical by mapping values.

3. Our Contribution

We present a privacy preserving scheme which involves the cryptographic type of privacy for the unrealized data sets constructed from the original data sets. The cryptographic method implemented is the RSA approach. Then the accuracy evaluation is performed with the help of Naïve Bayesian classification algorithm. The phases involved are as follows:

Data Set Pre-Processing

It is a very important step in data mining process. Data processing techniques, when applied before mining, can substantially improve the overall quality of patterns mined & the time necessary for actual mining. Data pre-processing techniques can improve the quality of the data, accuracy & efficiency of the mining process.

Unrealized Data Set Construction

To unrealize the samples, we start with both set of input sample data set and perturbing data set as empty sets. With respect to the procedure described above, universal data set is added as a parameter of the function because reusing pre-computed universal data set is more efficient than recalculating universal data set and this helps to save time of recalaculation. The recursive function unrealized training-set takes one data set in input sample data set in a recursion without any special requirement; it then updates perturbing data set and set of output training data sets correspondent with the next recursion[1].The original data set can be reconstructed if both of the unrealized data sets, i.e., both T^P and T' leaks out, so we go for cryptographic protection. The algorithm for unrealization of data sets is as follows[1].

```
INPUT   : $T_{s,}$ $T^U$ , $T^P$ , $T^'$
OUTPUT  : $T^P$ , $T^'$
1. We invoke the algorithm with $T^P$
   and $T^'$ as empty sets.
2. Take t which is a dataset in $T_s$
3. If t is not an element of $T^P$ or $T^P$
   ={t} then
4. $T^P = T^P + T^U$
5. $T^P =T^P -\{t\}$
6. t' is the most frequent dataset in
   $T^P$
7. Return the values ($T_s$–{t} , $T^U$ , $T^'$
   + {t'} , $T^P$ –{t'})
```

Cryptographic Privacy

Once the unrealized data sets are obtained the cryptographic approach of RSA is applied on the attributes of unrealized data set for enhanced protection and to prevent the leakage of unrealized data sets. The algorithm for cryptographic privacy is as follows[14].

```
INPUT   : ATTRIBUTE AGE (m) OF $T^P$ +$T^'$
OUTPUT  : ENCRYPTED ATTRIBUTE AGE (c)
1. Choose two distinct prime numbers p and q.
2. Find n such that n = pq.
3. Find the totient of n, $\phi(n)=(p-1)(q-1)$.
4. Choose e such that $1 < e < \phi(n)$
     – e and $\phi(n)$ share no divisors other than 1.
     – e is the public key which is used for encryption.
5. Determine d such that $de \equiv 1 \pmod{\phi(n)}$.
     – d is the public key used for decryption
6. Encryption : $c \equiv m^e \pmod{n}$.
7. Decryption : $m \equiv c^d \pmod{n}$.
```

Evaluation Using Naïve Bayesian

Naïve Bayesian[13] is a statistical classifier which performs probabilistic prediction, i.e., predicts class membership probabilities. It is based on Bayes theorem which is as follows:

$P(H|X)=[P(X|H)P(H)]/P(X)$

Let D denote a set of tuples and associated class labels. The tuple is denoted as X and there exists m classes $C_1,C_2\ldots C_m$.Applying Bayes theorem we get

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

The class with the largest value of $P(X|C_i)P(C_i)$ is assigned as the class value of the tuple.The accuracy is calculated using the following formula

$$\text{Accuracy} = \frac{\text{Correctly Classified Instances}}{\text{Total Number Of Instances}} \times 100$$

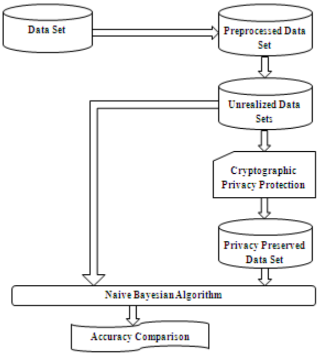

Fig. 1. System Architecture

4. Results And Discussions

The data set taken for evaluation is heart data set with 574 records and 7 attributes. The records with missing values are deleted as the preprocessing step, then the unrealization of data set is performed followed by the implementation of RSA over the attribute age for cryptographic protection in order to prevent the leakage of original data set from the unrealized data sets. Now the Naïve Bayesian classification algorithm is applied for evaluation of accuracy over the original data set, unrealized data sets and the cryptographic data set. The cryptographic privacy protected data set is able to give improved accuracy than the existing system.

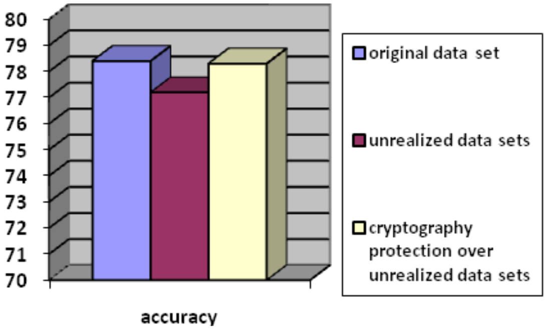

Fig. 2. Accuracy Comparison

5. Conclusion And Future Works

In this paper privacy preservation of data set is efficiently provided by the means of combination of unrealization of data sets and cryptographic approach. This work presents a new privacy preserving approach which removes each sample from a set of perturbing data set followed by a cryptographic approach with RSA. In future in order to improve the privacy preservation and the mining efficiency, an effective privacy preserving distributed mining algorithm of association rules can be proposed.

6.References

[1]Privacy Preserving Decision Tree Learning Using Unrealized Data Sets,Pui K. Fong and Jens H. Weber-Jahnke, Senior Member, IEEE Computer Society, IEEE Transactions on knowledge and data engineering, vol. 24, no. 2, 2012.

[2]Alexandre Evfimievski :Privacy – Preserving Data Mining by IBM Almaden Research Center, USA Tyrone Grandison IBM Almaden Research Center, USA.

[3] R. Agrawal and R. Srikant.2000, Privacy-preserving data mining. In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, TX, May 14-19 2000. ACM.

[4] Y. Lindell and B. Pinkas.2000, Privacy preserving data mining.In Advances in Cryptology – CRYPTO 2000, pages 36–54.Springer-Verlag.

[5] L. Sweeney,2002, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 571-588,

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkita-subramaniam.2006, l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24.

[7] Ninghui Li Tiancheng Li,Suresh Venkatasubramanian,t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.

[8] E.Poovammal and M. Ponnavaikko, 2009,Task Independent Privacy Preserving Data Mining on Medical Dataset International Conference on Advances in Computing,Control, and Telecommunication Technologies.

[9] Dinur and K. Nissim.,2003,"Revealing information while preserving privacy", PODS, pages 202–210.

[10] M. Kantarcioglu, J. Jin, and C. Clifton,2004, "When Do Data Mining Results Violate Privacy?" Proc Int'l Conf. Knowledge Discovery and Data Mining, pp. 599-604, 2004

[11] Agrawal D., Aggarwal C.,2002, "On the Design and Quantification of Privacy- Preserving Data MiningAlgorithms", ACM PODS Conference.

[12] R.Agrawal, J.kiernan, R.Srikant, Y. Xu,2002, "Hippocratic databases", 28th International Conference on very large Databases, Hong Kong, China.

[13] Jiawei Han and Micheline Kamber , 2006,"Data Mining: Concepts and Techniques", 2nd ed, ISBN 1-55860-901-6.

[14]http://en.wikipedia.org/wiki/RSA_(algorithm)