Preservation of the Privacy of Collected Data Samples without Reducing Utility of the Training Samples

Vana Surendra Kumar
M.Tech(Computer Science and Engineering),
Department of Computer Science and Engineering,
Sarada Institute of Science Technology and
Mangement, Srikakulam.

Jayanthi Rao Madina
Head of The Department,
Department of Computer Science and Engineering,
Sarada Institute of Science Technology and
Mangement, Srikakulam.

Abstract:

Privacy-preserving is an important issue in the areas of data mining and security. The aim of privacy preserving in data mining is to develop algorithms to modify the original dataset 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.

We proposed an efficient privacy preserving technique during the classification of data. We introduce a Cryptographic based approach that protects centralized sample data sets utilized for decision tree mining of data.

Preservation of privacy is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or reveal. In contrast to other sanitization approaches, our approach does not affect the accuracy efficiency of results of data mining .

The decision tree can be built directly from the pre- processed data sets, it means originals do not need to be formed. Moreover, this approach provides an efficient privacy preserving technique over data mining and can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected.

Keywords:

Classification, Data mining, Machine learning, Privacy protection, Cryptography and Security.

Introduction:

The problem of privacy-preserving in data mining has become more important in recent year because the ability to store personal data about users is increased, and the increasing knowledge about the data mining algorithms to control this information. There are number of techniques such as randomization and k-anonymity have been suggested in order to perform privacy-preserving data mining. Also this problem has been discussed in many communities such as the statistical disclosure control community, database community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar. This paper will try to explore different topics from the perspective of different communities, and will try to give a fused idea of the work in different communities.

Preserving privacy is more important for machine learning and data mining, but the measures designed to protect private information sometimes result in a degradation and reduced utility of the training samples. This work introduces an approach that can be applied to decision-tree learning, without concurrent loss of accuracy. It describes an privacy preservation approach for the collected data samples in cases when information of the sample database has been partially lost. This approach converts the original datasets into a group of unreal datasets, in which the original data cannot be reconstructed without the entire group of unreal datasets if some portion of the unreal datasets is stolen. This approach does not suitable when sample datasets have low frequency or low variance in the distribution of all samples. However, this problem can be resolved through a alternative implementation of the approach introduced later in this work, by using some extra storage. The key directions in the field of privacy-preserving data mining are as follows:

A. Privacy-Preserving Data Publishing:

These techniques tend to study different transformation methods associated with privacy. These approaches include methods such as randomization, k-anonymity, and l-diversity. A related issue is how the perturbed data can be used along with classical data mining methods such as association rule mining. Other related problems include that of determining privacy preserving methods to keep the underlying data useful or the problem of studying the various privacy definitions, and how they compare in terms of effectiveness in different states.

B. Modifying the results of Data Mining Applications to preserve privacy:

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has generate a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy. Likewise many techniques are available to modify the results of the data mining applications.

C. Cryptographic techniques for Distributed Privacy:

In many cases, the data may be distributed across many sites, and the owners of the data across these different sites may wish to compute a common function. In those cases, a variety of cryptographic protocols may be used to communicate among various sites, so that secure function computation is possible without revealing the sensitive information.

RESEARCH BACKGROUND AND OBJECTIVE :

Even when databases of samples with sensitive information are protected securely, partial information of the databases can be lost through procedural mistakes or privacy attacks which can be from anywhere within a network. This work focuses on analyzing privacy preservation following the loss of some training datasets from the whole sample database used for decision-tree learning.n this basis, we make the following assumptions for the scope of this work: first, as is the norm in data collection processes, a large number of sample datasets have been collected to achieve significant data mining results that cover the whole research target. Second, the number of datasets lost constitutes a small portion of the entire sample database.

Third, for decision-tree data mining, no attribute is designed for distinctive values, because such values negatively affect decision classification. The objective of this work is to introduce a new privacy preserving approach to the protection of sample datasets that are utilized for decision-tree data mining. Privacy preservation is applied directly to the samples in storage, so that privacy can be safeguarded even if the data storage were to be threatened by unauthorized parties. Although effective against privacy attacks by any unauthorized party, this approach does not affect the accuracy of data mining results. Moreover, this technique can be applied at any time during the data collection process, so that the protection of privacy can be in effect as early as the first sample is collected.

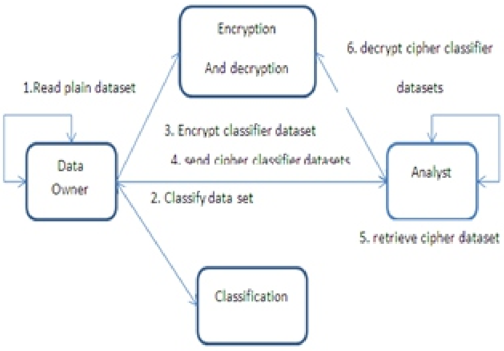
Existing System:

Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is to preserve customer privacy by perturbing the data values . In this scheme random noise data is introduced to distort sensitive values, and the distribution of the random data is used to generate a new data distribution which is close to the original data Distribution without revealing the original data values.

The estimated original data distribution is used to reconstruct the data, and data mining techniques, such as classifiers and Association rules are applied to the reconstructed data set and after refinement of this approach have tightened estimation of original values based on the distorted data. The data distortion approach has also been applied to Boolean values in research work.

Proposed System:

Now a day's providing privacy to unrealized data sets is very important issue. In this paper we proposed new concepts contain both feature of cryptography and classification. By implementing those concepts we can provide more security of unrealized datasets and also perform the classification of Synthetic dataset.


Encryption
And decryption
6. decrypt cipher classifier
datasets
1.Read plain dataset
3. Encrypt classifier dataset
4. send cipher classifier datasets
Data
Owner
Analyst
2. Classify data set
5. retrieve cipher dataset
Classification

Initialize Training Datasets:

Datasets are the collection of tuples with respect to different attributes and possible values for each attribute and with class labels, is given for the classification process for analyzing the testing set behaviour with machine learning approach. Synthetic dataset can be gathered for the classification of results. Initially data set can be forwarded to the encoder, encoder returns the cipher dataset.

Creation of Unrealized datasets:

Usually data can be passed to the analysts for the machine learning purpose, but there is a privacy preserving issue regarding the confidential information. So in this paper we introduced Tiny Encryption Algorithm for the privacy issue. After applying this mechanism dataset can be constructed as unrealized dataset. i.e cipher dataset can be passed to the analyst for the classification instead of plain sensitive or confidential information.

Classification of dataset using pruning rule based decision tree :

Pruning rule based decision tree is collection of algorithms for performing classifications in machine learning and data mining. It develops the classification model as a decision tree. pruning rule based decision tree consists of three groups of algorithm:pruning rule based decision tree using pruning rule based decision tree -no-pruning andpruning rule based decision tree -rules. In this summary, we will focus on the basic pruning rule based decision tree algorithm

Algorithm:

Pruning rule based decision tree is implemented recursively with this following sequence

1. Check if algorithm satisfies termination criteria

2. Computer information-theoretic criteria for all attributes

3. Choose best attribute according to the information-theoretic criteria

4. Create a decision node based on the best attribute in step 3

5. Induce (i.e. split) the dataset based on newly created decision node in step 4

6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call)

7. Attach the tree obtained in step 6 to the decision node in step 4

8. Return tree

For alternative description, check out Pruning rule based decision tree in pseudocode flavor below:

```
Input : An attribute value Dataset D

 Tree ={}

if D is pure or other stopping criteria met them

terminate
 end if

for all attribute a € D do
```

```
Compute information theoretic criteria if we split on a
End if
abest  = best attribute according above computed criteria
Tree=create decision node that test abest   in the root
Dv=Induced sub dataset  from D base on abest
For all Dv do
  Tree= Pruning rule based decision tree  (Dv)
Attach tree to the coreesponding branch of Tree
End for
Return Tree
```

Retrieval of Original classified results:

After generating the classification results, results can be passed to the Data owner, there administrator can perform attribute oriented decryption for the resulted set. Original data set can reconstruct by the decoder and classified rules can be obtained finally at the data owner end.

Comparative Analysis:

Classification is process of grouping together documents or data that have similar properties or are related. Our understanding of the data and documents become greater and easier once they are classified. We can also infer logic based on the classification. Most of all it makes the new data to be sorted easily and retrieval faster with better results. Recent proposal of privacy preserving during classification in data mining, mostly works on two approaches those are perturbation and randomization-based approaches and Cryptographic approaches, During the initial approach we inject fake values in to real dataset and converts into unrealized dataset. n the Cryptographic approach we convert the Plain data to cipher by using an cryptographic approach. The main drawback with the Previous approach is data retrievability, after retrieving the classified data from the analyst and the rules which are classified may not be loptimal due to imputation of the fake values in the real dataset, maintain the details of fake imputation rules for entire dataset(Both training and testing datasets) is a time consuming process, Our proposed approach provides more security from the third parties but obviously computation complexity depends on the number of records in the datasets obviously. For optimal security we are considering our cryptographic approach with AES.

Conclusion:

The privacy preserving process sometimes reduces the utility of training datasets, which causes inaccurate data mining results. Privacy preservation approaches focus on different areas of a data mining process, and data mining methods also vary. This paper focuses on privacy protection of the training samples applied for decision tree data mining. In this paper we proposed an efficient privacy preservation technique during classification of unreal datasets.

It prevents the data owner from the un authorized access and privacy issues, Our proposed approach works efficiently with our violating the classification properties. Meanwhile, an accurate decision tree can be built directly from those unreal data sets. Finally the results yield accurate results even though classification applies on the cipher dataset.

REFERENCES:

[1] Pui K.Fong and JensH.Weber-Jahnke, "Privacy preserving Decision tree Learning Using Unrealized Data sets", IEEE Trans. Knowledge and Data Eng., vol.24 No. 2, Feb 2012.

[2] S.Ajmani, R.Morris, and B.Liskov, "A Trusted Third-Party Computation Service,"Technical Report MIT-LCS-TR-847, MIT 2001.

[3] S.L.Wang and A.Jafari, "Hiding Sensitive Predictive Association Rules," Proc.IEEE Int'l Conf. Systems, Man and Cybernetics, pp.164-169, 2005.

[4] R.Agrawal and R.Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data (SIGMOD'00), pp.439-450, May 2000.

[5] Q.Ma and P.Deng, "Secure Multi-Party Protocols for Privacy Preserving Data Mining," Proc. Third Int'l Conf.Wireless Algorithms, Systems, and Applications (WASA'08), pp.526-537, 2008.

[6] J.Gitanjali, J.Indumathi, N.C.Iyengar, and N.Sriman, "A Pristine Clean Cabalistic Foruity Strategize Based approach for Incre-mental Data Stream Privacy Preserving Data Mining," Proc. IEEE Second Int'l Advance Computing Conf.(IACC), pp.410-415,2010.

[7] N.Lomas, "Data on 84,000 United Kingdom Prisonersis Lost,"Retrieved Sept.12, 2008, http://news.cnet.com/8301-1009_3-10024550-83.html, Aug.2008.

[8] BBC News Brown Apologizes for Records Loss. Retrieved Sept. 12, 2008, http://news.bbc.co.uk/2/hi/uk_news/politics/7104945.stm,Nov.2007.

[9] D.Kaplan, Hackers Steal 22,000 Social Security Numbers from Univ. of Missouri Database, Retrieved Sept.2008,http://www.scmaga-zineus.com/Hackers-steal-22000-Social-Security-numbers-from-Univ.-of-Missouridatabase/article/34964/May2007.

[10] D.Goodin,"Hackers In filtrate TD Ameritrade client Database, "Retrieved Sept.2008,http://www.channelregister.co.uk/2007/09/15/ameritrade_database_burgled/,Sept.2007.

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

[12] Y.Zhu, L.Huang, W.Yang ,D.Li, Y.Luo, and F.Dong, "Three New Approaches to Privacy-Preserving Add to Multiply Protocol and Its Application,"Proc.Second Int'l Workshop Knowledge Discovery and Data Mining, (WKDD'09),pp.554-558,2009.

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

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

[15] C.Aggarwal and P.Yu, Privacy-Preserving Data Mining:, Models and Algorithms. Springer, 2008.

[16] L.Sweeney, "k-Anonymity: A Model for Protecting Privacy,"Int'l J.Uncertainty, Fuzziness and Knowledge – basedSystems, vol.10, pp.557-570, May 2002.

[17] J.Dowd, S.Xu, and W.Zhang, "Privacy-preserving Decision Tree Mining Based on Random Substations," Proc. Int'lConf. Emerging Trends in Information and Comm. Security (ETRICS'06), pp.145-159, 2006.

[18] Bu, L.Lakshmanan, R.Ng, and G.Ramesh, "Preservation of Patterns and Input-Output Privacy," Proc. IEEE 23rd Int'lConf Data Eng., pp.696-705, Apr.2007.

[19] S.Russell and N.Peter, Artificial Intelligence. A Modern Approach2/ E. Prentice-Hall, 2002.

[20] P.K.Fong, "Privacy Preservation for Training Data Sets in Database: Application to Decision Tree Learning, master'sthesis, Dept.of Computer Science, Univ.of Victoria, 2008.

BIOGRAPHIES:

Vana Surendra Kumar

student in M.Tech(Computer Science and Engineering) in Sarada Institute of Science Technology and Management, Srikakulam. He has received his B-TECH(CSE)Sarada Institute of Science Technology and Mangement, Srikakulam.. His interesting areas are Data Mining, Networking and cloud computing.

Madina Jayanthi Rao

working as a HOD of CSE in Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh. He is pursuing Ph.d at KRISHNA UNIVERSITY Machilipatnam in computer science. He received his M.Tech (CSE) from Aditya Institute of Technology And Management (AITAM), Tekkali. Andhra Pradesh. His interest research areas are Data mining, Image Processing, Computer Networks, Distributed Systems. He published 12 international journals and he was attended number of conferences and workshops.