Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Restoration and Privacy Preserving of Data Using C4.5 Algorithm

Sharmila Harale*
Department of Computer Engineering, Pune University India

A. K. Bongale
Department of Computer Engineering, Pune University India

Abstract*— Data mining extracts knowledge to support a variety of areas as marketing, medical diagnosis, weather forecasting, national security etc successfully. There is an improved advance in hardware technology which increases the capability to store and record personal data about consumers and individuals. So there is a challenge to extract certain kinds of data without violating the data owners' privacy. As data mining becomes more enveloping, such privacy or security concerns are increasing. This gives birth to a new branch of data mining method called privacy preserving data mining algorithm (PPDM). The aim of this algorithm is to protect the easily affected information in data from the large amount of data set. The privacy preservation of data set can be expressed in the form of decision tree. This paper proposes a privacy preservation based on data set complement algorithms which store the information of the real dataset. So that the private data can be safe from the unauthorized party, if some portion of the data can be lost, then we can recreate the original data set from the unrealized dataset and the perturbed data set and also this approach can deal with both discrete and continuous data for which continuous data is converted into discrete data which is the basis of this paper.

Keywords*— Data mining, Privacy Preserving Data Mining (PPDM), Decision Tree, Decision Tree Learning, ID3 algorithm, C4.5 algorithm, Unrealized Dataset, Data Perturbation, Dataset Complementation, Discrete and Continuous data.

I. INTRODUCTION

Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, and anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the coming up need for turning such data into useful information and knowledge for many broad applications including market analysis, business management, and decision support, data mining has fascinated a great deal of interest in information industry in recent years. Data collected from information providers are important for pattern reorganization and decision making. The data collection process takes time and efforts hence sample datasets are sometime stored for reuse. However attacks are attempted to take these sample datasets and private information may be leaked from these stolen datasets. Therefore privacy preserving data mining algorithms are developed to convert sensitive datasets into sanitized version r altered version in which private or sensitive information is hidden from unauthorized or unofficial retrievers.

Privacy Preserving Data Mining (PPDM) refers to the area of data mining that aims to protect sensitive information from illegal or unwanted disclosure. Privacy Preservation Data Mining was introduced to preserve the privacy during mining process to enable conventional data mining technique. Many privacy preservation approaches were developed to protect private information of sample dataset.

Modern research in privacy preserving data mining mainly falls into one of two categories: 1) perturbation and randomization-based approaches, and 2) secure multiparty computation (SMC)-based approaches. SMC approaches employ cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed among different parties and they take part in the information computation and communication process. SMC research focuses on protocol development for protecting privacy among the involved parties or computation efficiency; however, centralized processing of samples and storage privacy is out of the scope of SMC [1].

We introduce a new perturbation and randomization based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to alter or sanitize the samples earlier to their release to third parties in order to moderate the threat of their accidental disclosure or theft. In contrast to other sanitization methods, our approach does not affect the accuracy of data mining results. The decision tree can be built directly from the altered or sanitized data sets, such that the originals do not need to be reconstructed. In addition to this, this approach 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 [1].

A. Decision Tree Learning

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Decision tree learning is a method for approximating discrete valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be represented as sets of if-then rules to improve human readability.

II. LITERATURE SURVEY

a) P. K. Fong et al. in [1] introduces a privacy preserving approach that can be applied to decision tree learning, without related loss of accuracy. It describes an approach to the protection of the privacy of collected data samples in cases where information from the sample database has been partly lost. This approach converts the original sample data sets into a group of altered or unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. In the meantime, an accurate decision tree can be built directly from those unreal data sets. This new approach can be applied directly to the data storage as soon as the first sample is collected. The approach is well-matched with other privacy preserving approaches, such as cryptography, for extra protection.

b) J. Dowd et al. in [2] , proposed the several contributions towards privacy-preserving decision tree mining. The most important is that the framework introduced a new data perturbation technique based on random substitutions. This perturbation technique is similar to the randomization techniques used in the context of statistical disclosure control but is based on a different privacy measure called ρ1-to-ρ2 privacy breaching and a special type of perturbation matrix called the γ-diagonal matrix.

c) Aggarwal et al. in [3], Privacy Preserving Data Mining: Models and Algorithms Aggarwal and Yu categorize privacy preserving data mining techniques, including data modification and cryptographic, statistical, query auditing and perturbation-based strategies. Data modification techniques maintain privacy by modifying attribute values of the sample data sets. Essentially, data sets are modified by eliminating or unifying uncommon elements among all data sets. These similar data sets act as masks for the others within the group because they cannot be distinguished from the others; every data set is loosely linked with a certain number of information providers.

d) L. Sweeney et al. in [7], K-anonymity: A Model for Protecting Privacy is a data modification approach that aims to protect private information of the samples by generalizing attributes. K-anonymity trades privacy for utility. Further, this approach can be applied only after the entire data collection process has been completed.

e) L. Liu et al. in [4] proposed a new method that we build data mining models directly from the perturbed data without trying to solve the general data distribution reconstruction as an intermediate step. More precisely, proposed a modified C4.5 decision tree classifier that can deal with perturbed numeric continuous attributes. Privacy preserving decision tree C4.5 (PPDTC4.5) classifier uses perturbed training data, and builds a decision tree model, which could be used to classify the original or perturbed data sets. The experiments have shown that PPDTC4.5 classifier can obtain a high degree of accuracy when used to classify the original data set.

III. IMPLEMENTATION DETAILS

In Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is to preserve customer privacy by disturbing the data values. In this method random noise data is introduced to alter 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.

The other approach uses cryptographic tools to construct data mining models. The goal is to securely build an ID3 decision tree where the training set is distributed between two parties. Different solutions were given to address different data mining problems using cryptographic techniques. ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree.

A. Disadvantage of Existing System

Existing system covers the application of new privacy preserving approach with the ID3 decision tree learning algorithm and the limitation of this algorithm is, it works for discrete-valued attributes only.

B. Proposed Solution

One of the limitations of ID3 decision tree algorithm can be overcome by using C4.5 algorithm, an ID3 extension and data mining methods with mixed discretely and continuously valued attributes and thus data restoration and preservation of privacy of data is done.

ID3's sensitivity to features with large numbers of values is illustrated by Social Security numbers. Since Social Security numbers are unique for every individual, testing on its value will always yield low conditional entropy values. However, this is not a useful test. To overcome this problem, C4.5 uses a metric called "information gain," which is defined by subtracting conditional entropy from the base entropy; that is, Gain (P|X) =E (P)-E (P|X). This computation does not, in

itself, produce anything new. However, it allows you to measure a gain ratio. Gain ratio, defined as Gain Ratio (P|X) =Gain (P|X)/E(X), where E*(X)* is the entropy of the examples relative only to the attribute. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too-much details in the training data set. Like IDE3 the data sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute (Quinlan, 1993). Decision trees are built in C4.5 by using a set of training data or data sets as in ID3. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sub lists [13].

The following assumptions are made for the scope of this paper: first, as is the norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target. But as considering security of data, here small amount of sample dataset from real dataset is taken. Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth, all data collected are discretized; continuous values can be represented via ranged value attributes for decision tree data mining [1].

C. Mathematical Model

Initial definition of ID3 is restricted in dealing with discrete sets of values. It handles symbolic attribute effectively. However, discrete sets of values have to convert to continuous-valued attributes (numeric attribute) to fit the real world scenario.

1) Converting continues valued attributes to discrete valued attributes

To convert continuous valued attributes to discrete, a new discrete valued attributes that partition the continuous valued attribute into symbolic attribute again is defined. For an attribute A which has numeric values, a new boolean value that is true when A<=c and false otherwise. The only thing is to compute the best threshold c.

In an example, the most information gain is attribute 'outlook'. In the subset rooted at 'outlook: sunny', need to compute the information gain for 'Humidity' which is a numeric attribute. To do this, sorting is done in ascending order of values of 'Humidity' attribute.

Humidity	0.68	0.72	0.87	0.9	0.91
Play	Yes	Yes	No	No	No

The aim is to pick a threshold that produces the greatest information gain. By sorting the numeric attribute values, then identifying adjacent examples that differ in their target classification, a mean of the values of these attributes is calculated, Humidity> (0.72+0.87)/2 that is Humidity>0.795, a set of candidate threshold is generated. Then compute information gain for each candidate and find the best one for splitting.

2) Dataset Complementation Approach

In this segment work is done with the sets that can contain multiple instances of the same element. The segment begins by defining fundamental concepts and then data unrealization algorithm.

a) T - Data Table

b) T_S - Training Set, is constructed by inserting sample data sets into a data table.

c) T^U - Universal set of data table T is a set containing a single instance of all possible data sets in data table T.

d) T^P – Perturbed Data Set.

e) T' – Unrealized Training Set

3) Algorithm for Data Unrealization

Dataset Complementation approach was designed for discrete value classification so continuous values are replaced with ranged values. The entire original dataset is replaced by unreal dataset for preserving the privacy via dataset complementation. This approach 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. The original accuracy of training dataset is preserved without linking the perturbed dataset to the information provider i.e. accurate data mining results are yields while preserving privacy of individual's records by dataset complementation approach.

A data complementation approach requires an extra table T^P for converting sample dataset Ts into an unrealized training set T'. T^P is perturbing set that generates unreal dataset.

Initially T' and T^P are an empty set. When we get an Ts the T^P is constructed with universal set T^U by adding T^U into T^P. Whenever we get sample data item t in T_S we remove it from T^P and transfer one data item ti^1 into T^1. Ti^1 is the latest available frequent data item in T^P. When traversing T^P is finished and if sample data item t^1 is not available in T^P then again add universal set T^U into T^P.

To unrealized the samples T_S, initialize both T' and TP as an empty sets, i.e. invoke the above algorithm with Unrealized_Training_set(Ts, TP, { }, { }) . The elements in the resulting data sets are unreal individually, but meaningful when they are used together to calculate the information required by a modified C4.5 algorithm [13].

4) C4.5 Algorithm for Decision Tree Generation based on Information Entropy and Information Gain

The algorithm C4.5 selects a test attribute (with the smallest entropy) according to the information content of the training set T_s. The information entropy and information gain functions are given as below.

Information entropy is a term that was introduced by Claude Shannon's information theory in 1948. In information theory, information content is measured in bits. Entropy measures the minimum number of bits necessary to communicate information. It can also be used to measure the uncertainty associated with a random variable. If a random variable *X* has possible outcomes *ki* with probabilities ( ) *i P k* while *i* is an integer and 1 V *i* V *n*, then the information content *I* in bits can be expressed by:

$H(X) = I(P(k_1), P(k_2)\ldots..,P(k_n) = -\sum_{i=1}^{n} P(k_i) log_2\ P(k_i)$ [4]

Information content *I* indicate the uncertainties of event *X*. *I* is ranged from 0 to– $log_2\ (\frac{1}{n})$, while 0 means the event is absolutely biased and– $log_2\ (\frac{1}{n})$, means the event is fair.

Information gain measures the gain in information content by a classification event of an attribute test. If *T* is the training set, *a* is the test attribute with possible values *i k* (*i* is an integer and 1 V *i* V *n*) and *d* is the decision attribute with possible values *i v* (*j* is an integer and 1 V *j* V *m*), then information gain *Gain* is shown as following:

Gain (*a*) = $H_d\ (T) -\ H_d\ (T/_a)$ where $H_d\ (T)$ is the information content of *d* before the test, equals:

The higher the information gains of an attribute test, the lower the uncertainty contained in its decision. Therefore, by comparing the information gain among the attributes available as an internal node, one can find the best test attributes in the decision-tree learning process [1].

Pseudo Code of C4.5:

1. Check for base cases.
2. For each attribute a calculate:
   i. Normalized information gain from splitting on attribute a.
3. Select the best a, attribute that has highest information gain.
4. Create a decision node that splits on best of a, as root node.
5. Recurs on the sub lists obtained by splitting on best of a and add those nodes as children node [12].

5) Modified C4.5Decision Tree Generation Algorithm and Dataset Reconstruction

As entropies of the original data sets, T_s, can be determined by the retrievable information—the contents of unrealized training set, T', and perturbing set, T^p—the decision tree of T_s can be generated by Modified C4.5Decision Tree Generation Algorithm .

IV. System Architecture

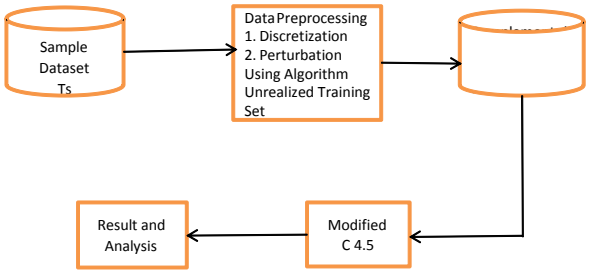

Fig. 1 System Architecture

The system architecture of proposed system is shown in above figure. It consist of two main module first is data preprocessing and second is decision tree generation. In data preprocessing module initially continuous value attribute dataset is converted into discrete value after that the dataset is converted into sanitized version by using algorithm unrealized training set. Then generated complemented dataset and perturbed dataset is given as an input to decision tree generation module in which decision tree is built by using C4.5 and result generated by algorithm is compared to analyze the algorithm.

Software Requirements contain front end as JDK 1.5 and above versions, tool as Net beans, operating system as windows 7 and back end as MySQL server.

Hardware Requirements contain RAM minimum 1 GB, hard disk of minimum 20 GB and other standard input output devices.

V. Actual Input Dataset and Actual Result Set

Outlook	Humidity	Windy	Play
Sunny	0.9	Weak	No
Sunny	0.87	Strong	No
Overcast	0.93	Weak	Yes
Rainy	0.89	Weak	Yes
Rainy	0.8	Weak	Yes
Rainy	0.59	Strong	No
Overcast	0.77	Strong	Yes
Sunny	0.91	Weak	No
Sunny	0.68	Weak	Yes
Rainy	0.84	Weak	Yes
Sunny	0.72	Strong	Yes
Overcast	0.94	Strong	Yes
Overcast	0.74	Weak	Yes
Rainy	0.86	Strong	No

Fig. 2a Sample Dataset

Outlook	Humidity	Windy	Play
Sunny	High	Weak	No
Sunny	High	Strong	No
Overcast	High	Weak	Yes
Rainy	High	Weak	Yes
Rainy	Normal	Weak	Yes
Rainy	Normal	Strong	No
Overcast	Normal	Strong	Yes
Sunny	High	Weak	No
Sunny	Normal	Weak	Yes
Rainy	Normal	Weak	Yes
Sunny	Normal	Strong	Yes
Overcast	High	Strong	Yes
Overcast	Normal	Weak	Yes
Rainy	High	Strong	No

Fig. 2b Discrete Dataset

Outlook	Humidity	Windy	Play
Sunny	1.038	Strong	Yes
Sunny	0.440	Weak	Yes
Sunny	0.572	Strong	No
Sunny	0.153	Strong	Yes
Overcast	1.595	Weak	No
Overcast	1.773	Strong	No
Overcast	1.775	Strong	Yes
Overcast	0.809	Weak	No
Overcast	0.186	Weak	Yes
Overcast	0.057	Strong	No
Rainy	1.640	Weak	No
Rainy	1.208	Strong	No
Rainy	1.773	Strong	Yes
Rainy	0.610	Weak	No

Fig. 2c Perturbed Dataset

In this project, the actual input dataset (Fig. 2a) is weather dataset having 14 records of any real world dataset containing both discrete and continuous values have been collected to achieve significant data mining results covering the whole research target.

The actual input dataset gets converted to discrete dataset as shown in Fig. 2b having all discrete values for the continuous values of input dataset.

The result set is perturbed dataset (Fig. 2c) contain sanitized data of given input dataset which can be safely published as a data for mining without disclosure of any correct records thereby maintain the privacy of the original dataset owner.

IV. CONCLUSION

This paper covers the system architecture and mathematical model of decision tree classifier for preserving privacy via dataset complementation. The entire original dataset is first converted to discrete dataset and then replaced by using algorithm Unrealize_training_set. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. During the privacy preserving process, this set of perturbed datasets is dynamically modified. As the sanitized version of the original samples, these perturbed datasets are stored to enable a modified decision tree data mining method.

Privacy preservation via data set complementation fails if all training data sets are leaked because the data set reconstruction algorithm is generic. Therefore, further research is required to overcome this limitation. As it is very simple to apply a cryptographic privacy preserving approach, such as the (anti)monotone framework, along with data set complementation, this direction for future research could correct the above limitation.

This paper covers the application of this new privacy preserving approach with the C4.5 decision tree learning algorithm and for both discrete-valued attributes and continues –valued attributes and also data can be restored from perturbed Tp and T' (Tdash) dataset if some portion of original dataset is lost. But to recover original dataset, the whole datasets Tp and T' are required, on the portion of the Tp and T' , original dataset cannot be obtained.

REFERENCES

[1] Pui K. Fong And Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets" Proc. IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 2, February 2012.

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

[3] C. Aggarwal and P. Yu, Privacy-Preserving Data Mining:, Models and Algorithms. Springer, 2008.
[4] L. 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.
[5] Y. Lindell and B. Pinkas "Privacy preserving data mining" In Advances in Cryptology, volume 1880 of Lecture Notes in Computer Science, pages 36–53. Springer-Verlag, 2000.
[6] P.K. Fong, "Privacy Preservation for Training Data Sets in Database: Application to Decision Tree Learning," master's thesis, Dept. of Computer Science, Univ. of Victoria, 2008.
[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, pp. 557-570, May 2002.
[8] S. Ajmani, R. Morris, and B. Liskov, "A Trusted Third-Party Computation Service," Technical Report MIT-LCS-TR-847, MIT, 2001.
[9] S.L. Wang and A. Jafari, "Hiding Sensitive Predictive Association Rules," Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 164- 169, 2005.
[10] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data (SIGMOD '00), pp. 439-450, May 2000.
[11] 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.
[12] Surbhi Hardikar, Ankur Shrivastava, Vijay Choudhary,"Comparison Between ID3 And C4.5 In Contrast To IDS", Proc. VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667.
[13] Tejaswini Pawar1 , Prof. Snehal Kamlapur,"Decision Tree Classifier for Privacy Preservation", Proc. IJETCAS 12-391, 2013.
[14] Payam Emami Khoonsari and AhmadReza Motie," A Comparison of Efficiency and Robustness of ID3 and C4.5 Algorithms Using Dynamic Test
and Training Data Sets", Proc. International Journal of Machine Learning and Computing, Vol. 2, No. 5, October 2012.

