RESEARCH PAPER | Engineering | Volume : 4 | Issue : 3 | Mar 2014 | ISSN - 2249-555X

A Novel Privacy Preserving Approach for Decision Tree Learning

KEYWORDS: ID3, C4.5

Dr. S.Vijayarani
Assistant Professor, School of Computer Science and Engineering, Bharathiar University, Coimbatore

M.Sangeetha
M.Phil Research Scholar, School of Computer Science and Engineering, Bharathiar University, Coimbatore

ABSTRACT *Data mining is the extraction of the hidden information from large databases. Preserving privacy against data mining algorithms is a new research area. The problem of privacy preservation in data mining has become more important in recent years because of increasing need to store vast data about users. In this research work, a new privacy preserving approach is applied to decision tree learning. The original sample datasets are converted into a group of unreal datasets. An accurate decision tree is built using those unreal datasets. Many decision tree learning algorithms are used to generate decision tree such as CART, CHAID, and Ripper. In this research work, decision tree learning algorithms namely ID3 and C4.5 algorithms are used for building decision tree. A new modified decision learning approach is proposed for generating an accurate decision tree. The performance of the decision tree learning algorithms and the proposed technique are evaluated.*

1. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases and also a powerful new technology with great potential to analyze important information in their data warehouses.It is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. Data mining is the method of extracting patterns from data. It can be used to uncover patterns in data but is often carried out only on sample of data. The mining process will be ineffective if the samples are not good representation of the larger body of the data. The discovery of a particular pattern in a particular set of data does not necessarily mean that pattern is found elsewhere in the larger data from which that sample was drawn.

Many privacy protection approaches preserve private information of sample datasets, but not precision of data mining outcomes. This paper provides an approach that preserves privacy and utility of sample datasets for decision-tree data mining. In this research work, a new privacy preserving approach is applied to decision tree learning. The original sample datasets are converted into a group of unreal datasets. An accurate decision tree is built using those unreal datasets.

A new "perturbation and randomization based approach" and "k-anonymity approach" is applied in this paper. It protects centralized sample data sets utilized for decision tree data mining. The decision tree can be built directly from the sanitized data sets, such that the originals need not to be reconstructed.

This paper is organized as follows: Section 2 explains a brief discussion about the decision tree learning. Section 3 provides discussion on the previous works related to the topic. Section 4 describes the existing approaches of decision tree learning and the proposed algorithm for decision tree learning. Section 5 involves the Conclusion and future works.

2. DECISION TREE LEARNING

Decision tree learning uses a decision tree as a predictive model. The goal is to create a model that predicts the value of a target variable based on several input variables. Decision tree learning is one of the most widely used and practical methods for inductive inference. There are many specific decision-tree algorithms. Notable ones include:

ID3 (Iterative Dichotomiser 3)
C4.5 (successor of ID3)

CART (Classification And Regression Tree)
CHAID (CHi-squared Automatic Interaction Detector)

3. RELATED WORKS

A wide research has been devoted to the protection of sensitive information when samples are given to third parties for processing or computing [2], [3], [4], [5], [6]. Samples may be leaked or stolen anytime during the storing process or while residing in storage. This paper focuses on preventing such attacks to the samples by third parties.

Contemporary 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 [7].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 [8] for protecting privacy among the involved parties [9] or computation efficiency [10]; however, centralized processing of samples and storage privacy is out of the scope of SMC.

This paper extends the research work [11] and proposes a novel technique for privacy preserving decision tree learning by building decision tree using a new modified algorithm. The main advantage of using this modified algorithm is that it built decision tree using information gain and does not involve time complexity.

4. PROBLEM DEFINITION & PROPOSED METHODOLOGY

Privacy preservation decision tree learning is important concept in data mining. Decision tree should be built efficiently. In this paper we have proposed a privacy preserving approach that can be applied to decision tree learning. This approach converts the original sample datasets into a group of unreal datasets. The original sample datasets cannot be reconstructed from it. Meanwhile, an accurate decision tree is built from those unreal datasets.

- Unrealized dataset conversion
- Decision Tree Generation
- Distribution
- Comparison

Volume : 4 | Issue : 3 | Mar 2014 | ISSN - 2249-555X

UNREALIZED DATASET CONVERSION:
For conversion of unrealized dataset, we use the algorithm of unrealized training set. Modification module is the process of modifying the original sample dataset into the unreal datasets. Data modification techniques maintain privacy by modifying attribute values of the sample data sets. For this process K-anonymity approach is used for the modification purpose. K-anonymity is a data modification approach that aims to protect private information of the samples by generalizing attributes. In this process datasets are inserted into the data table. Data unrealization algorithm is used for this process. Inserted dataset are unreal dataset. Every data set in the data table is loosely linked with a certain number of information providers.

First we load the universal set and the sample set. This sample set and universal set is implemented by the unrealized training set algorithm. Finally the output of the unrealized data set is training set and perturbation set.

T^U, the universal set of data table T, is a set containing all possible datasets in data table T. Let T associates with attributes<Wind, Play> where Wind= {Strong, Weak} and Play= {Yes, No} then T^U= {<Strong, Yes>, <Strong, No>, <Weak, Yes>, <Weak, No>}.T_S is constructed by inserting sample data sets into a data table. T^P is a perturbing set that generates unreal datasets which is used for converting T_S into unrealized training set T'.

Algorithm Unrealize-Training-Set (T_S,T^U, T' , T^P)
Input: T_S , a set of input sample data sets
T^U , a universal set
T' , a set of output training data sets
T^P, a perturbing set

Output: T' , T^P
1. if T_S is empty then return(T', T^P)
2. t ← a data set 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' ←the most frequent dataset in T^P
7. return Unrealize-TrainingSet
(T_s-{t},T^u,T'+{t'},T^P-(t'))

DECISION TREE GENERATION:
Decision tree process has the process of providing the decision tree for the original dataset. So information provider only understands that information about the particular dataset.

ID3 Algorithm
The well-known ID3 algorithm builds a decision tree by calling algorithm Choose-Attribute recursively. This algorithm selects a test attribute (with the smallest entropy) according to the information content of the training set T_s.

Algorithm Generate-Tree (T_S , attribs , default)
Input: T_s, the set of training data sets
attribs, set of attributes
default, default value for the goal predicate
Output: tree, a decision tree

1. if T_S is empty then return default
2. default ←Majority _ Value(T_S)
3. if $H_{ai}(T_S)$ = 0 then return default
4. else if attribs is empty then return default
5. else
6. best ←Choose-Attribute(attribs, T_S)
7. tree← a new decision tree with root attribute best
8. for each value v_i of best do
9. T_{Si}← {datasets in T_s as best = k_i}
10. subtree ←Generate-Tree(T_{Si}, attribs-best, default}
11. connect tree and subtree with a branch labelled k_i
12. return tree

SYSTEM DESIGN

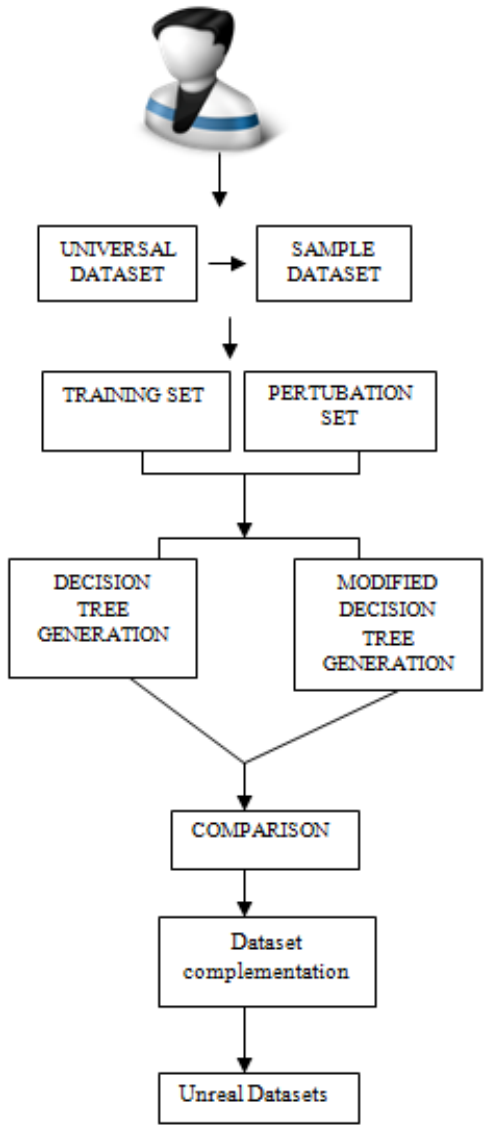


Figure 1: System Architecture of Proposed Methodology

C4.5 Algorithm
1. Check for base cases.

2. For each attribute a ,
- Find the normalized information gain from splitting on a.

3. Let a_best be the attribute with the highest normalized information gain.

4. Create a decision node that splits on a_best.

5. Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node.

Modified Decision Tree Algorithm
Input: sample data set

Output: Decision Tree

1. Load input data set for training .

2. If attribute is uniquely identify in data set , remove from it.

Volume : 4 | Issue : 3 | Mar 2014 | ISSN - 2249-555X

3. On the basis of distance metric divide the given training data in to subsets.

4. Calculate the distance for (1……N) each instance in available dataset

5. if (distance>55% && <70%) then instance is belong to same group and ,add in to new set and remove from original data set. Otherwise do nothing .

6. Repeat the step 4 and 5 for each instance until all matched is not found

7. On each subset apply ID3 algorithm recursively.

- If at target attribute all example are positive then return single tree root with label positive.
- If at target attribute all example are positive then return single tree root with label negative.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Otherwise
  - o Calculate the entropy of decision node , if entropy is not equal to zero than calculate information gain for each attribute.
  - o For spitting each choose that value whose information gain is maximum.
  - o Apply algorithm is recursively until entropy is not reaches to the zero of each attribute.

End

DATASET RECONSTRUCTION

Modified decision tree learning algorithm generates decision tree by using the unrealized training set, T′, and the perturbing set, T^p. Alternatively, have reconstructed the original sample data sets, T_S, from T′ and T^p, followed by an application of the proposed algorithm for generating the decision tree from T_S. The reconstruction process is dependent upon the full information of T′ and T^p. The reconstruction of parts of T_S based on parts T′ and T^p is not possible.

COMPARISON GRAPH:

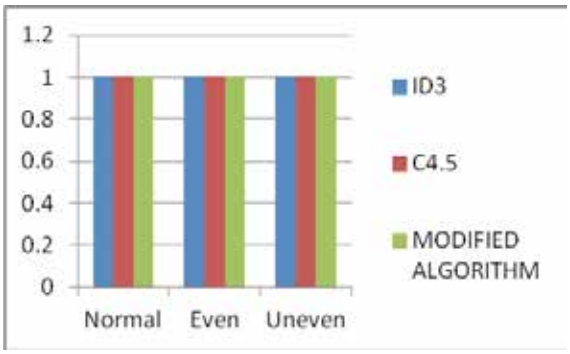
The experimental results from the application of the data set complementation approach to

1. normally distributed samples
2. evenly distributed samples
3. extremely unevenly distributed samples
4. randomly picked samples, where
   (i) was generated without creating any dummy attrib ute values and
   (ii) was generated by applying the dummy attribute technique to double the size of the sample do main.

For the samples (1-3), we will study the output accuracy (the similarity between the decision tree generated by the regular method and by the new approach), the storage complexity (the space required to store the unrealized samples based on the size of the original samples) and the privacy risk (the maximum, minimum, and average privacy loss if one unrealized data set is leaked).

Accuracy Graph

The decision tree generated from the unrealized samples (by algorithm Generate-Tree′) is the same as the decision tree(s), generated from the original sample T_S by the existing methods when samples are in normal distribution, even distribution and also in uneven distribution.

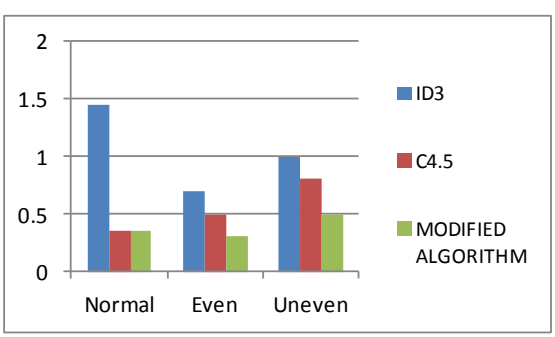


Storage Complexity Graph

The storage complexity is an effective performance measure that estimates the amount of space needed for storage that is required by ID3, C4.5 and the proposed modified algorithm for building an efficient and accurate decision tree.

The storage requirement for the data set complementation approach increases, while the required storage may be doubled if dummy attribute values technique is applied to double the sample domain. The best case happens when samples are evenly distributed, as the storage requirement is the same as for the originals. Samples with even distribution are taken. In even distribution, all datasets have the same counts. Decision tree is generated with increased storage required in existing method while it is not with the proposed method.

The worst case happens when the samples are in uneven distribution. Based on the randomly picked tests, Time complexity of storage for the proposed approach is less than five times (without using dummy values) and eight times (with dummy values) than that of the original samples.



5. CONCLUSION

In this research work, a new privacy preserving approach via data set complementation which confirms the utility of training data sets for decision tree learning has been proposed. This approach converts the sample data sets, training set, into some unreal data sets, such that any original data set is not able to reconstruct, if an unauthorized party where to steal some portion of (unrealized training sets and perturbing training set). Meanwhile, there remains only a low probability of random matching of any original data set to the stolen data sets, leaking data. This work covers the application of this new privacy preserving approach with the C4.5 and ID3 algorithms and discrete-valued attributes only. Thus it shows the outperformance of the proposed approach for preserving the privacy in decision tree learning. Finally the proposed approach is evaluated with the existing approaches based on the accuracy in generating the decision tree and result shows that the proposed approach performs better.

Future research should develop the application scope for other algorithms, such as C5.0, and data mining methods with mixed discretely—and continuously valued attributes.

Volume : 4 | Issue : 3 | Mar 2014 | ISSN - 2249-555X

REFERENCE

[1] Arun K Pujari: Data Mining Techniques, Universities Press (India) Private Limited 2001. | | [2] S. Ajmani, R. Morris, and B. Liskov, "A Trusted Third-Party Computation Service," Technical Report MIT-LCS-TR-847, MIT, 2001. | [3] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data (SIGMOD '00), pp. 439-450, May 2000. | [4] 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. | [5] S.L. Wang and A. Jafari, "Hiding Sensitive Predictive Association Rules," Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 164- 169, 2005. | [6] J. Gitanjali, J. Indumathi, N.C. Iyengar, and N. Sriman, "A Pristine Clean Cabalistic Foruity Strategize Based Approach for Incremental Data Stream Privacy Preserving Data Mining," Proc. IEEE Second Int'l Advance Computing Conf. (IACC), pp. 410-415, 2010. | [7] 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. | | [8] 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. | | [9] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02), pp. 23- 26, July 2002. | | [10] Dr. S.VIJAYARANI, Ms. M.SANGEETHA, "An Efficient Technique for Privacy Preserving Decision Tree Learning", INDIAN JOURNAL OF APPLIED RESEARCH (IJAR) – ISSN – 2249-555X, Volume 3, Issue 9, September 2013, P.No. 127-130. |