Efficient Decision Tree Generation with Privacy Using Perturbation

[1]NencyGhetia, [2]Prof. Nitin J. Rola,
[1]PG student, [2]Assistant professor,
[2]Department of computer engineering
[1]Darshan Institute Of Engineering and Technology, Rajkot, India

***Abstract* - In recent years, advances in hardware technology have led to an increase in the capability to store and record personal data about consumers and individuals. This has led to concerns that the personal data may be misused for a variety of purposes. Privacy-preserving is an important issue in the areas of data mining and security. The aim of privacy preserving 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. The data set complementation approach expands the sample storage size (in the worst case, the storage size equals (2|TU-1|*|TS|) ; perturbation will improve some storage size using like c5.0 algorithm. We will optimize the processing time when generating a decision tree from those samples and functional dependencies. This paper work on optimize processing time, improve storage size, reduce processing time and functional dependencies.**

***Index Terms* - classification, data mining, security, cryptography**

I. INTRODUCTION

Data mining is used for retrieve knowledge or some pattern from big data.This is mainly used by researchers for their research in any of their work. Privacy preservation data mining is special technique for protect sensitive data and keep it private.

Privacy preserving is more important topic now a days .personal data of users is increased day by day, so we have to increase knowledge. The knowledge data mining algorithms to control it. There are many techniques for privacy preserving data mining. This problem is discussed in many communities like cryptography community and database community. This paper will try to expand some different topics related to data mining and related communities. Privacy preserving is important for machine learning and data mining[2] , but measures designed to protect private information sometimes degrade performance and less utility of samples. This approach applied to decision tree learning , without loss of accuracy. This approach is used for collected data samples in which information is partially lost of data samples.This approach converts the original data set into unreal data set[1], in which original data set can't reconstructed if entire unreal data set is not available. This approach is not suitable for sample data set which have low frequency or low variance. This problem can be resolved with some alternative approach introduce in this paper.

WhereverTimesisspecified,TimesRomanorTimesNewRomanmaybeused.Ifneitherisavailableonyourwordprocessor,pleaseusethefontclosestinappearancetoTimes.Avoidusingbit-mappedfonts.TrueType1orOpenTypefontsare required.Pleaseembed all fonts, in particular symbolfonts,aswell,formath,etc.

Analysis

In privacy preserving data mining models and algorithms include different techniques like cryptography techniques, purterbationbased,statistical, query auditing.

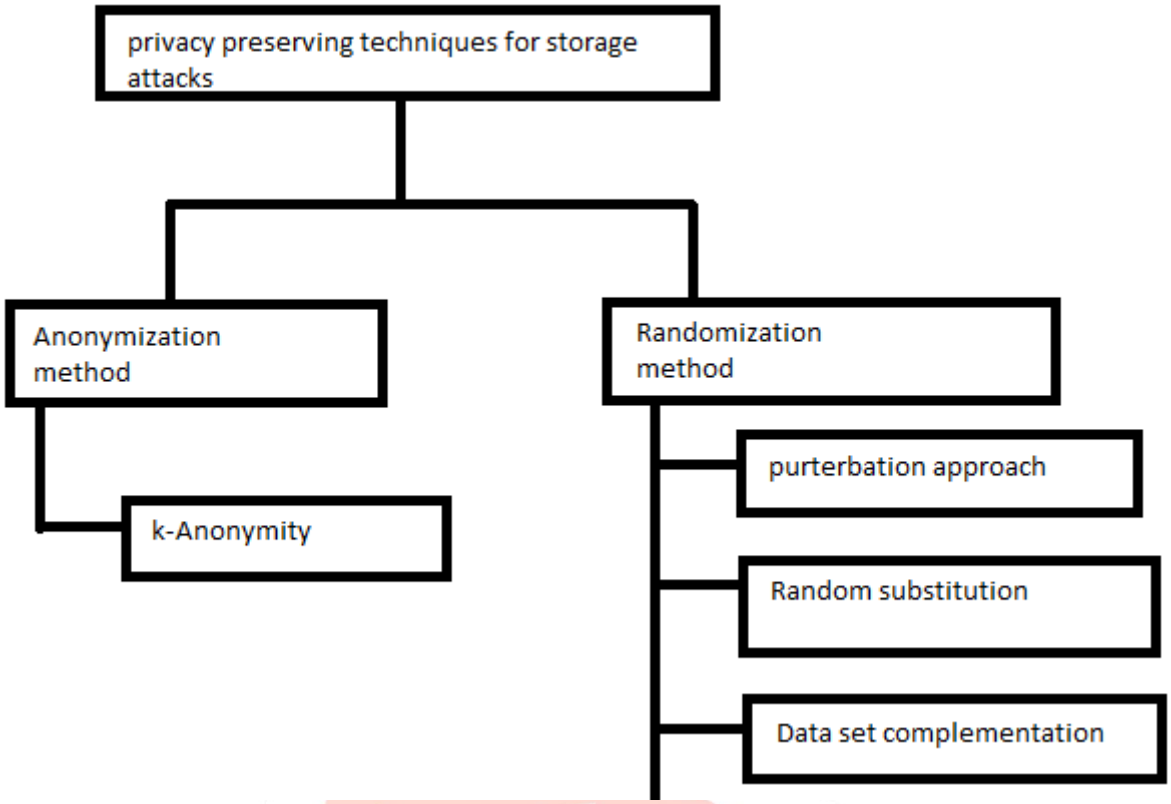

Fig 1

- Data publishing - This type is related with transformation techniques like randomization [4], k-anonimity.other issues is how purebred data along with association rule mining [3], how to determine privacy preserving methods to keep the data useful.
- Randomization method - In this attribute set value is masked by adding noise or replace with random value.
- Perturbationapproach - In this protect privacy of data through distortion information from original data set.
- Data set complementation approach - This is data perturbed approach in which original data set to unreal data set.
- Decision tree classifier - Decision tree is one of the important methods used for numerical data to classify with less computation. Each non-leaf node called internal node or splitting node contains a decision and most appropriate target value assigned to one class is represented by leaf node [5].

II. Objective and Background

Database samples are protected properly even if some information is lost through some mistake or attacks.our work focuses on privacy preserving techniques following the loss of some data set from whole data set for that we have use decision tree learning.[6,7].

We take some assumptions: A large number of datasets have been collected. Some data sets are lost from large data sets. No attribute is designed for distinctive values because its negatively affect in classification of decision tree.

Privacy preservation is important for machine learning and data mining, it mainly used for designed protect private information often result in a trade-off: reduced utility of the training samples.In existing system they introduced a new privacy preserving approach with use of data set complementation which confirms the utility of training data sets for decision tree learning. This approach converts the original data set in some unreal data set. Now this new approach use perturbation for privacy preservation. This approach will improve storage size support continuous attribute also.

III. Existing System

Existing system use ID3 algorithm. Existing system describesan 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 [1], 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.

Existing system have unrealized data set complexity is O(Ts). They make one universal set for input of the function. So result is, matching rate is one third of the unprotected samples of best case. Sanitization process complexity is O(Ts). So in worst case storage needed for unprotected sample requires is 2*|Tu|-1. So proposed system we have to reduce storage size of the data
Set.

Limitations in existing system are insufficient storage mechanism and this ID3 only can be implemented for discrete-valued attributes. This system support continuous variable.These issues will overcome in this paper.

IV. PLAN FOR PROPOSED SYSTEM

In this we will use improved algorithms which is more efficient than c4.5 or other related data mining algorithms. This algorithm is used for decision tree. flow of work is as follows:

Flow:

Input dataset -----> storage of dataset in database -----> preprocessing of data -------> purterbation added to main data ------> make unrealized data set -------> use c5.0 for decision tree construction -----> results

Advantages of improved algorithm rather than other algorithm like ID3 or C4.5

- Speed
- Small decision tree
- Boosting facility
- Less memory
- Winnowing
- Less Misclassification cost

V. CONCLUSION

This paper shows different privacy preservation technique, applied to c5.0 algorithms. This research will improve storage size of unreal data sets and functional dependency. Future work improve100% functional dependency.

VI. ACKNOWLEDGMENT

We wish to avail the opportunity to express our sincere gratitude to our institute and second author Prof Nitin J. Rola to support me and thankful them for my work and make them successful.

REFERENCES

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

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

[5] TejaswiniPawar*, Prof. SnehalKamalapur "A Survey on Privacy Preserving Decision Tree Classifier", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 6, November- December 2012, pp.843-847

[6] 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.

[7] 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

[8] Fong, P. K., and Jahnke, J. H. W. , "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets" ." IEEE Transl. on knowledge anddata engineering, vol. 24, no. 2, February2012.

[9] Aggrwal C. C..,Philip S Yu., " Privacy preserving data mining models and Algorithms.", Springer Science+Business media.,LLC..2008.

[10] M. R. Pawar,MampiBhowmik ,"Privacy preserving decision tree learning using unrealized data sets", IJREAS Volume 3, Issue 3 (March 2013)