An Efficient un-realization algorithm for privacy preserving decision tree learning using McDiarmid's bound

T. Satyanarayana Murthy, N.P.Gopalan, D.Yakobu

Abstract— JESD204B transmitter is a part of the serialized data interface between logic devices and data converters bases on the JESD204B standards. The organization, where this project is currently executed, is currently developing the JESD204B Tx and Rx IP for the avionics spacecraft applications where the fail proof and function safe and reliable data communication is essential. The verification of this IP is an important phase in the development wherein it is extremely important to perform rigorous tests on the design to confirm its acclaimed functionality and performance. The verification of an IP of this complexity is done in a systematic and efficient way using the Universal Verification Methodology which is basically constructed using the SystemVerilog. A verification environment is built using the UVM to verify the functionality of the IP. The test cases are written to verify each functionality of the design and the randomized stimuli are applied to cover all the possible input scenarios. The code coverage and the functional coverage is determined and further stimuli are applied to achieve the target coverage. The verification of the JESD204B Transmitter IP is completed with a functional coverage of around 39.17% for each test instance and an overall functional coverage of 100% and code coverage of 94.25%. The verification environment can be reused with minor changes to verify the JESD204B Receiver IP.

Key words: Coverage, IP verification, JESD204B, Universal Verification Methodology.

I.INTRODUCTION

The privacy preservation, with the introduction of privacy preserving data mining (PPDM) by agrawal in the year 2006 [1,2,3,4,5] became a important and evolving topic. These techniques use decision trees for privacy preservation. These decision trees are simple, but a powerful form of multiple variable analysis. It uses heuristic measures such as Gini index, information gain for classifying the attributes. Earlier hoeffding inequalities are used based on Hoeffding bound for construction of decision tree.It suffers with

1. The Hoeffding inequality is not sufficient tool and the results obtained are not absolutely true.
2. The McDiarmid's inequality an alternative way to achieve privacy using modified McDiarmid's algorithm.

II. LITERATURE STUDY

Un-realization Algorithm has been used in data perturbation, data modification and data swapping. Perturbation and randomization approaches were discussed by Aggarwal and Yu P.S [6] . k-Anonymity [7] a data modification technique uses modified attribute values to achieve privacy. Random substitutions which changes the value frequently is an advanced perturbation approach[8]. Data Swapping [9] technique swap values in dataset by using a t-order statistics model . Probability distribution for data distortion was implemented by Chong k.Liew et al., [10]. The important operations for data distortion are, identifying density function, reconstruction of the original series and generating a disturbing series. Perturbation technique discussed by Rakesh Agrawal et al., uses Gaussian distribution, noise adding and random substitution, cannot handle periodic updates on the data, and also cannot be applicable to categorical attributes. Privacy-preserving scalar product protocol for data mining was developed by Jaideep Vaidya et al., [11]. SMC algorithm proposed by Lindell et al., [12] used for data hiding. Privacy was achieved as the algorithm contains ID3. Data perturbation approach was discussed by Hilol Kargupta at al.,[13] which was based on random matrix based filtering. C. Aggarwal et al.,[14] developed a condensation approach for hiding the sensitive data. Machanavajjhala et al., [15] developed an l-diversity technique which overcome the drawbacks of k-anonymity for data Anonymization. Handling continuous sensitive attributes is the drawback of l-diversity. To overcome all these attacks and limitations decision tree approaches are implemented by Pui K. Fong et al.,[16,17] based on dataset complementation approach. Decision tree algorithms likeID3 algorithm and modified ID3 algorithm were used to achieve the privacy but these cannot solve periodic updates that occur the dataset. As it uses universal set approach,Storage complexity increases . Rutkowski et al., [18] discussed a decision trees based on the mcdiarmid's bound for data streams. An application of Hoeffding's inequality for decision trees construction using data streams by Piotr Duda et al., [19] . T.Satyanarayana Murthy et al.,[23,24,25,26,27] stated meta heuristic based algorithms to improve association rule hiding. In this paper proposed two algorithms which gives better outputs than traditional algorithms.

III. DATASET COMPLEMENTATION APPROACH

Definition 1. Z^U, the state space set.

$$|Z^U| = y_1 * y_2 * \ldots *y_x, \quad (1)$$

Definition 2. $|pZ^U| = p*|Z^U|$. (2)

Revised Manuscript Received on December 22, 2018.

T. Satyanarayana Murthy, National Institute of Technology, Tiruchchirapalli, Tamil Nadu, INDIA (e-mail : murthyteki@gmail.com)
N.P.Gopalan, National Institute of Technology, Tiruchchirapalli, Tamil Nadu, INDIA (e-mail : npgopalan@nitt.edu)
D.Yakobu, Vignan's Foundation for Science, Technology and Research, Guntur, Tamil Nadu, INDIA (e-mail : yakobdsp83@gmail.com)

IV. ARCHITECTURE OF UNREALIZED ALGORITHM ALONG WITH MCDIARMID'S BOUND

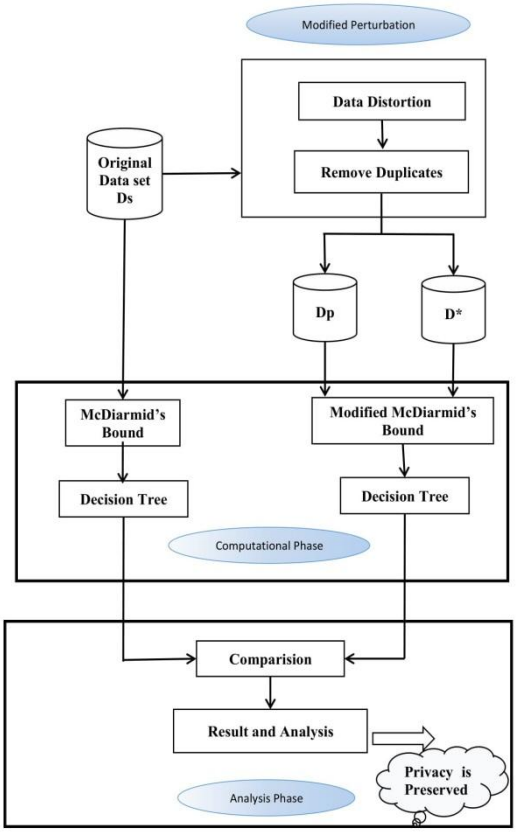

Figure 1. Architecture of Unrealized algorithm along with McDiarmid's Bound

The above figure shows the architecture where it consists of three different phases. The first phase is the Distortion phase, here the original data set Ds is subjected to distortion through modified algorithm. In this phase all the duplicate data elements will be removed and generates a Perturbed data set Dp and an un-realized data set Du. The second phase is the Computational phase, here McDiarmid's Bound is applied on original data set Ds that results a decision tree and Modified McDiarmid's Bound is applied on the Perturbed and un-realized data sets Dp, Du respectively, which also results a decision tree. The third phase is the Analysis phase, where both the generated decision trees are compared and analyzed for Privacy Preservation.

V. MODIFIED UNREALIZED ALGORITHM

The modified un-realization algorithm removes the duplicates from the perturbed dataset, so that the size of the perturbed dataset is small as compared to the un-realization algorithm.

Modified Algorithm:

Modified_Unrealized_Algorithm(Ds,Pu,Du,Dp)

Input : {DS, Pu, Du, Dp }
Output:{ Du, Dp }

Modified_Unrealized Algorithm(Ds,Pu,Du,Dp){
if(Ds_Sample_Dataset==null)
{
return{ Du_Unrealized_set,Dp_perturbing_set }
}
T1=a data element of Ds_Sample_Dataset
if(T1.isElementOf(Dp_perturbing_set){
Dp_perturbing_set= Dp_perturbing_set-T1;
T11=a data element of Dp_perturbing set;
if(T11 .Exists_Multiple) {
T111=T11; }}
else{
Dp_perturbing_set=Dp_perturbing_set+ Du_unrealized_set -T1;
T11=a data element of Dp_perturbing set
if(T11 .Exists_Multiple)
{
T111=T11;
}}
return Unrealized_Algorithm(DS_Sample_Dataset- T1, Pu_State Space,Du_Unrealized set+ T11,Dp_perturbing set-T11- T111)
return {Du_Unrealized_set,Dp_perturbing_set }}

V. MCDIARMID'S BOUND

As decision trees evolved, they turned out to have many useful features, both in the traditional fields of science and engineering and in a range of applied areas and different methods including such as Game theory, Artificial Intelligence, Machine Learning, Soft computing, general statistics, Image and Signal Processing, Database systems, regression analysis. These trees produce results that communicate very well in symbolic and visual terms. There are Id3,c4.5, CART, and CHAID which are based on the decision tree induction. Decision trees representation is rich enough to represent any discrete-value classifier. Decision trees are capable of handling data sets that may have missing values. Decision trees are considered to be a nonparametric method. They are capable of handling errors in datasets. The decision trees have assumptions about the space distribution and the classifier structure. ID3 (Iterative Dichotomiser) was the first decision tree algorithm. Later classification and Regression Trees and C4.5 algorithms has developed for privacy preservation. These were used to solve problems distinguished by static data. Most of the algorithms (like ID3 and C4.5) require that the target attribute will have only discrete values. McDiarmid's bound based decision trees are robust.

McDiarmid's Bound for Gini Index

In this paper, we consider Gini gain of McDiarmid's Bound rather than information gain since ε tends to zero for Gini gain. The McDiarmid's Bound for Gini index:

$$Gini(A) = 1 - \sum_{i=1}^{j} (\frac{b^i}{B})^2$$

Where b_i be the number of elements in set A and B be total number of values in entire data set.

$$Gini_{xi}(A) = \frac{bi}{B}(1 - \sum_{i=1}^{j} (\frac{b^i}{B})^2)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

where i be number of attributes i.e. 1,2,3…j.

Gini x (A) = min{Gini xi (A)} where i be number of attributes i.e. 1,2,3…j.

Gini Gain notation is as follows:

ΔGini x(A)=Gini(A)-Gini x(A)

The calculations that are obtained Using McDiarmid's Bound are as follows:

Gini (Ds) =0.489

Gini Age (Ds)=0.228

ΔGini (Age) = Gini (Ds) – Gini Age (Ds)= 0.489 – 0.228 =0.261.

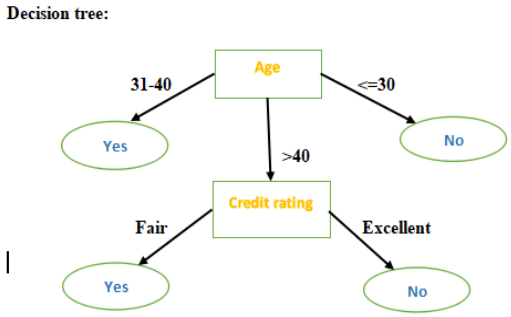

Figure 2. Decision Tree of Training Data Set (Ds)

VI. MODIFIED MCDIARMID'S BOUND

Modified McDiarmid's Bound uses Gini Index as d*ΔGini (D) , sample data set Ds (ΔGini (Age) = 0.261) and (Du + Dp) (d*ΔGini (Age) = 0.290) get nearly same values . So, generated decision tree will also be same and data privacy is preserved.

Modified McDiarmid's Tree Algorithm

Inputs: Q set of examples,
W collection of attributes,
E( ) splitting function, D*Gini, Gain function,
α reducing probability parameter a choosing node.

Output: McDT A decision Tree generated from McDiarmid's Algorithm

Procedure McDiarmidTree(Q, W, E, α)
Let McDT single(node).m1 then leaf node.
Let Wm1 = Ws
For each class Nr
For each attribute w ϵ W
For each value wλ of attribute w
Let $O^r_{w\lambda}$ (m1) = 0.
For each example A in B
Sort A using McDT.
For each attribute w ϵ Wm
For each value wλ of attribute w
If value of example A for attribute w equals wλ and A comes from class r
Increment $O^r_{w\lambda}$ (m).
Label l with the majority class.
If the examples seen so far at l are not of the same class, then
Compute $\bar{E}_m(w)$ for all the attributes using $O^r_{w\lambda}$ (m).
Compute w_{MAX1} ,w_{MAX2}.
Compute ε.
If g*$\bar{E}$m(w_{MAX1}) - g*$\bar{E}$m(w_{MAX2}) > ε, then
Replace m by an internal node that splits on w_{MAX1}.
For each branch of the split
Add a new leaf mp, and let
Wm_p = Wm \ {w_{MAX1}} at mp.
For each attribute S_r, w ϵ W, w_λ of w
Let $O^r_{w\lambda}$ (mp) = 0.
Return McDT.

The calculations that are obtained Using Modified McDiarmid's Bound are as follows:

E=Dp +Du

Gini (E) =0.497

Gini $_{Age}$ (E)= 0.468

ΔGini (Age) = Gini (Ds) – Gini $_{Age}$ (Ds)= 0.497 – 0.468 =0.029.

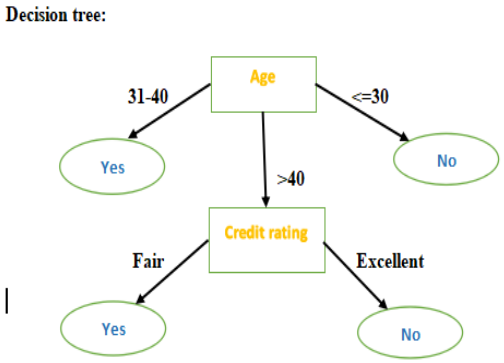

Figure 3 Decision Tree of Training Data Set (E)

VII. RESULT ANALYSIS

Experiments are performed on using python program and WEKA software on i3 processor with 8GB RAM. Python program accepts the original data and process it and produces results. The initial values of Un-realized algorithm<Ds,Pu,Du,Dp> are <7,12,0,0>.Initially Ds contains 7 record, Pu contains 12 record, Du and Dp contains empty record. The result produced by the Un-realized algorithm is <0,12,7,10>.Modified Un-realized algorithm produces <0,12,7,6> as compared with the traditional algorithm Dp values is reduced.

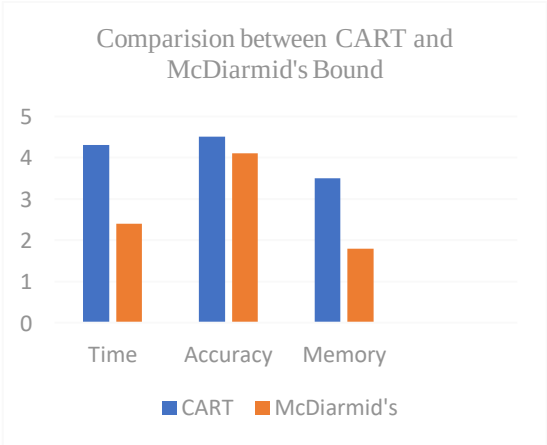

Figure 4. Comparison between CART and McDiarmid's Bound

By experimental results between cart and McDiarmid's bound, we get above graph on considering processing time, accuracy and memory as parameters . In the above figure 5 the accuracy is nearly same for both algorithms but the remaining parameters makes McDiarmid's bound to be efficient than cart. McDiarmid's bound takes less processing because of its linear function of B while CART takes power function of B. The size of training data set also effects the cart but not on McDiarmid's Bound.

VIII. CONCLUSION

In this digital world, information has been retrieved from many social networking sites like whatsapp, instagram and snapchat etc. this collected information has to keep very confidential and couldn't be used for any kinds of uses. Several techniques/algorithms are used to hide or protect the collected data by adding noise to the original data and generate a distorted dataset. In this paper modified McDiarmid's bound has been proposed based on information gain and gini index. The result of McDiarmid's bound is quiet efficient than traditional algorithms. Further instead of McDiarmids approach advanced classifiers has been use to improve the performance.

REFERENCES

1. Rakesh Agrawal, Tomasz Imielinski, Arun Swami," Mining association rules between sets of items in large databases". ACM SIGMOD international conference on Management of data SIGMOD, pp. 207,1993.
2. Agrawal,R.,Srikant,R., a. Quest Synthetic Data Generator . IBM AlmadenRe- searchCenter ⟨http://www.Almaden.ibm.com/cs/quest/syndata.html⟩,1994.
3. Chen, M.S.,Han,J.,Yu,P.S.,"Datamining:An overview from a database per spective". IEEE Trans. Knowl. DataEng.8(6),866–883,1996.
4. Aggarwal,C.C.,Pei,J.,Zhang,B.,."On privacy preservation against adversarial data mining",ACM SIGKDD International Conference on Knowledge Dis- covery and Data Mining,pp.510–516, 2006.
5. T.Satyanarayana Murthy, "Pine Apple Expert System Using Improved C4.5 Algorithm", pp-1264-1266,(2013).
6. Aggarwal C.C., Yu P.S, Privacy preserving Data Mining:, Models and Algorithms.Springer,2008.
7. Sweeney,L. : k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570,2002.
8. J. Dowd, S. Xu, W. Zhang, "Privacy-preserving decision tree mining based on random substitutions", Proc. Int. Conf. Emerg. Trends Inf. Commun. Security, pp. 145-159, 2006.
9. T. Dalenius and S.P. Reiss, "Data-Swapping: A Technique for Disclosure Control," Journal of Statistical Planning and Inference, vol. 6, pp. 73-85, 1982.
10. C. K.Liew ,U.J.Choi , C.J.Liew, "A Data distortion by probability distribution" ACM TODS ,pp 395-411,1985.
11. J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002, IEEE 2002.
12. Y. Lindell, B.Pinkas, "Privacy preserving data mining", in proceedings of Journal of Cryptology, 5(3), 2000.
13. H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On thePrivacy Preserving Properties of Random Data PerturbationTechniques", in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.
14. C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT), pp. 183–199, 2004.
    Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k-Anonymity", Proc. Int'l Con! Data Eng. (ICDE), p. 24, 2006.
15. Pui K. Fong and Jens H. Weber-Jahnke, Senior Member ,"Privacy Preserving Decision Tree Learning Using Unrealized Data Sets ", Proceeding of the IEEE Transactions On Knowledge And Data Engineering,VOL. 24, NO. 2, pp. 353 -364, FEB 2012.
16. P.K. Fong, J.H.Weber-Jahnke "Privacy preservation for training data sets in database: Application to decision tree learning, " M.SC Thesis,2012.
17. L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, "Decision trees for mining data streams based on the mcdiarmid's bound," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 6, pp. 1272-1279, 2013.
18. Piotr Duda, Maciej Jaworski, Lena Pietruczuk, Leszek Rutkowski, "A novel application of Hoeffding's inequality to decision trees construction for data streams", Neural Networks (IJCNN) 2014 International Joint Conference on, pp. 3324-3330, 2014.
19. N.P.Gopalan, T.Satyanarayana Murthy, Yalla Venkateswarlu, "Hiding Critical Transactions using Un-realization Approach", IJPAM, Vol 118,No.7 ,629-633,2018.
20. T.Satyanarayana Murthy, N.P.Gopalan, "A Novel Algorithm for Association Rule Hiding", International Journal of Information Engineering and Electronic Business (IJIEEB), Vol.10, No.3, pp. 45-50, 2018. DOI: 10.5815/ijieeb.2018.03.06
21. T.Satyanarayana Murthy, N.P.Gopalan, Yalla Venkateswarlu," An efficient method for hiding association rules with additional parameter metrics ",IJPAM ,Vol 118,No.7,285-290,2018.
22. T.Satyanarayana Murthy, N.P.Gopalan, , "Association rule hiding using chemical reaction optimization",SCOPRO 2017 Conference,IIT Bhubaneswar,2017,(Accepted).
23. T.Satyanarayana Murthy.,"Privacy Preserving for expertise data using K-anonymity technique to advise the farmers", International Journal of Electrical, Electronics and Data Communication, Volume-1, Issue-10, 2013.
24. Satyanarayana Murthy, T., & Gopalan , N.P., Sai Krishna .A, (2018), "The Power of Anonymization and Sensitive Knowledge Hiding Using Sanitization Approach", International Journal of Modern Education and Computer Science(IJMECS), Vol.10, No.9, pp. 26-32, 2018.DOI: 10.5815/ijmecs.2018.09.04.
25. Satyanarayana Murthy, T., & Gopalan , N.P., Sasidhar Gunturu.,(2018) " A Novel Optimization based Algorithm to Hide Sensitive Item-sets through Sanitization Approach", International Journal of Modern Education and Computer Science(IJMECS), Vol.10, No.10, pp. 48-55, 2018.DOI: 10.5815/ijmecs.2018.10.06.
26. T.V. Babu, T.S. Murthy, B. Sivaiah, "Detecting unusual customer consumption profiles in power distribution systems – APSPDCL", 2013 IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC), pp. 1-5, December 2013.