International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 12 (2017) pp. 3090-3096

Novel Algorithm for PPDM of Vertically Partitioned Data

Hemlata
Research Scholar,
Department of Computer Science and Applications,
Maharshi Dayanand University, Rohtak, Haryana, India.

ORCID: 0000-0002-6105-7399

Dr. Preeti Gulia
Assistant Professor,
Department of Computer Science and Applications,
Maharshi Dayanand University, Rohtak, Haryana, India.

ORCID: 0000-0001-8535-4016

Abstract

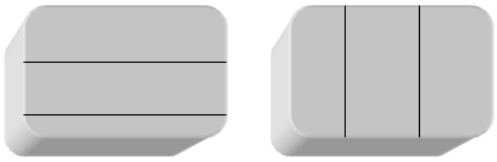
Privacy preserving data mining (PPDM) technique means mining or extracting knowledge from the data without sacrificing the sensitive or private data to the users. Protecting the private data from malicious users is an open challenge in the current scenario. So, everybody is not open to share the private data. Several methods of data mining have been proposed with the condition of concealing sensitive or personal data. In the paper, the methods are studied in detail and the comparison is presented in tabular form. Also, a new proposed PPDM algorithm is presented for building a decision tree.

Keywords: Data Mining, privacy preserving, decision tree, vertically partitioned data

INTRODUCTION

In today's technological world, data can be personal or organizational. Data Mining (DM) extract the knowledge or hidden meaning from the whole data which otherwise cannot be used. But it can sometimes reveal some sensitive or private data of some data owners. The owners want some protocol which can protect their sensitive data from being public. So analysts try to encrypt the original data with computed data which do not reveal the original data. This process is known as Privacy-Preserving Data Mining (PPDM). It is a biggest challenge in this data sensitive world.

Now-a-days, due to the increase in volume of data i.e. Big Data, the data is distributed among different servers or sites. The partitioning of data can be of two types- Horizontal partitioning and vertical partitioning. In horizontal partitioning the tuples of the data are divided among multiple sites. While in vertical partitioning different attributes are present on different sites i.e. only the data related to some attributes are present on one site.



Horizontal Partitioning Vertical Partitioning

It's a biggest challenge in today's scenario to preserve the private data in the distributed data which is partitioned. There is a lot research work in this area to propose the solution to safeguard the personal or private data.

The paper is organized as follows: Section II summarises the related work done in the area. Critical Analysis of the research done is recapitulated in tabular form in Section III. Section IV presents the new proposed algorithm to create a decision tree for vertically partitioned data. Future work related to the proposed algorithm is presented in Section V.

LITERATURE REVIEW

Many researchers have done work in the area of Privacy Preserving Data Mining. Existing work focuses on protecting private data and proposing new models of PPDM. Various methods of PPDM were discussed like-Anonymization, Association Rule, Classification, Secure Multi-party Computations, Perturbation techniques, Decision tree, 3-D Clustering, Cryptographic method and WHT etc. Different research works done in the area are summarized as follows:

Charu C. Aggarwal et.al [1], proposed the condensation approach of PPDM. In this method raw data was condensed to various groups of specific sizes and then statistics was used to analyse the data. The size of the group refers to the level of privacy-preserving. As the size of the group increases, the amount of privacy also increases. This is a simple privacy

preservation method but it has a limitation that it in this approach information is lost.

Nan Zhang et.al [2] introduced an Algebraic- technique-based scheme. This new scheme can give more accurate results as compared to randomization approach. Moreover this new scheme can be integrated with the existing systems as middleware. In the scheme, there is a very important criterion known as maximum acceptable disclosure level for the data provider. There are two basic components of the proposed method - the perturbation guidance (PG) component for data miner to calculate $k*$, the disclosure level and the perturbation guidance Vk*.

Animesh Tripathy et.al [3] used the classification method of PPDM in which secure multi-party computations were used. Also the differences between entropy and gini were demonstrated and concluded that high degree of accuracy and privacy are given by pruning. The training data of contact-lenses (24 rows and 5 columns) was used for analysis. The whole analysis was done in Weka which has cross validation test mode. The analysis showed that correct classified instances are given by pruned trees. With the decrease in the size of tree, correctness and accuracy increases.

Weiwei Fang et.al [4], proposed a new decision tree algorithm based on homomorphic encryption technology. The proposed architecture consists of databases, a miner,and a calculator. The new scheme used is additive homomorphic encryption and decryption scheme. It was concluded that the approach provides privacy preserving, accuracy and efficiency. H.R.Jalla and P.N. Girija [5], proposed an approach of PPDM for horizontal partitioned data which is based on linear transformations like WHT and perturbation techniques. Experiments were done on Iris and WDBC, the real datasets. Analysis was done for different combinations of linear transformations (WHT-WHT, WHT-DCT, WHT-FISIP and WHT-SAN) by using Weka tool. The results show that the proposed approach gave the accuracy similar like K-NN classifier.

Nasrin Irshad Hussain et.al [6], proposed a new privacy preserving data mining method for big data. Cryptography and clustering techniques were used. Rule based system was used for clustering. Different rules were choosen for different data-sets. New Accessment method was given having three layers- Secret Layer, Authorised Layer and Public Layer. Sumana M, et.al [7] presents a splitting strategy and a semi-trusted third party to create binary decision tree model. Faris Alqadah and Raj Bhatnagar[8] presented a novel algorithm to discover 3-Cluster in vertically partitioned data. Mathematical formulations were also presented to measure the quality of clusters. Rosa Karimi Adl et.al [9] modelled a game approach for finding consensual privacy protection levels by using anonymization method. The new approach demonstrated a sequential game as a process of private data collection. To examine the perfect symmetry of the game backward induction method was used.

Omar Abdel Wahab et-al [10] proposed a privacy-preserving Distributed Association Rule Mining approach-DARM for answering association rules queries in a distributed environment for achieving the objective of data privacy and query confidentiality. The aim of DARM model is to generate new rules by preserving original data. It contains three parties- data providers, data consumers, and master miner. Experimental results show that the approach is very efficient when the number of attributes in the query increases. Vikas Ashok and Ravi Mukkamala[11] suggested a new scheme in which the owner of the data derives association rules locally and sends them to the third party for extracting new data. This scheme reduces the generation of spurious rules. Sheng Zhong and Zhiqlang Yang [12] proposed a new PPDM Perturbation method – Guided Perturbation. This method gives the accuracy similar to cryptographic method and is much faster than Cryptographic method. Yi Xia et.al [13] proposed a new algorithm Recursive Estimation (RE). As compared to the conservative randomization factors, use of non-uniform randomization factors improves the accuracy. Yanguang Shen et.al [14] used semi-honest third party for collaborative computation by implementing PPDM decision tree algorithm-PPC4.5. Gopal Behra[15] proposes protocols based on secure multiparty computations for privacy preserving C4.5. The protocol do not require third party server. The results show that the proposed protocol of C4.5 gave better results as compared to ID3. F. Emekci* et.al [16] proposes a novel method to build a decision tree using ID3 algorithm. Also an efficient method was proposed to verify the correctness of the results when a large number of parties are involved. Weiwei Fang et.al [17] presents a novel privacy preserving decision tree learning method. Experimental results show that the method provides a good capability of preserving .privacy, accuracy and efficiency.

Saeed Samet et.al [18] presents a secure multi-party sub-protocol to show Gini can be used to create decision tree classifications. Saeed Samet et.al [19] presents new protocol both for back-propagation algorithm (BP) and extreme learning machine algorithm (ELM). The model is securely shared among all parties. Huafeng Ba et.al [20] presented two algorithms- IPFS and KIPFS to find the key features from the candidate semi-id feature set. Annonymizing the identified KIPFS, achieve better performance. Yanguang Shen et.al [21] gave an algorithm of privacy preserving C4.5.Secure scalar product protocol and secure sum protocol are used for collaborative computing to protect privacy effectively. . Hemlata Chahal [22] modified ID3 algorithm and implemented on real dataset. Jinfei Liu et.al [23] proposed two protocols for privacy preserving DBSCAN clustering. The protocols improved the privacy of partitioned data

CRITICAL ANALYSIS

There are many techniques proposed for PPDM. But different methods are used in different situations and with different datasets. No single method is apt for all datasets and all situations. A comparison of the work of different researchers has been presented in the tabular form (Table 1) in ascending order of the work done. The attributes taken of critical analysis are PPDM technique, Approach used, Results and Future scope of the work.

Table: Critical Analysis

S. No	Authors	Publication	PPDM Technique used	Approach	Result and Accuracy	Limitations/ Future Scope
1	Charu C. Aggarwal and Philip S. Yu	EDBT, 2004	Annonymization	Condensation approach	The proposed condensation framework led to privacy.	Leads to information loss.
2	Yi Xia, Yirong Yang and Yun Chi	ACM, 2004	Association Rule mining	New Algorithm-RE(Recursive Estimation)	Use of non-uniform randomization factors improves the accuracy.	In future, randomisation factor will be used to suit individual's privacy concerns.
3	Nan Zhang, Shenguan Wang and Wei Zhao	ACM, 2005	Classification	Algebraic-technique-based scheme	Accurate classifiers can be built and private information is not disclosed	In future, the scheme may be applied for clustering and cryptography techniques.
4	Sheng Zhong and Zhiqlang Yang	Springer, 2007	Perturbation Technique	Guided Perturbation method- a novel approach for accuracy and privacy	It gives accuracy similar to cryptographic method and is much faster than Cryptographic method.	The solution still has a tradeoff between the privacy preservation of the two involved parties.
5	F. Emekci *, O.D. Sahin, D. Agrawal, A. El Abbadi	Elsvier, 2007	Classification method- Decision Tree	ID3 algorithm	Modified algorithm was proposed to achieve privacy preserving.	Large dataset can be taken
6	Faris Alqadah and Raj Bhatnagar	ACM, 2008	3- Clustering technique	3-Cluster in vertical partitioned data and mathematical formulation to measure the quality.	The algorithm gives high quality clusters which are efficient in time.	Further alternate methods of clustering can be used for measurement.
7	Weiwei Fang and Bingru yang	IEEE, 2008	Classification method- Decision Tree	Decision Tree learning method	It provides good capability of preserving .privacy, accuracy and efficiency.	Extension of work can be done by using other methods like clustering and association rule, etc.
8	Saeed Samet and Ali Miri	IEEE, 2008	Classification method- Decision Tree	Gini index used for creating decision tree classification	The protocol was efficient and practical.	Extension of the proposed protocol by using vertically partitioned data.
9	Yanguang Shen, Hui Shao and Jianzhong Huang	IEEE, 2009	Privacy preserving C4.5	Collaborative Computing by secure scalar	The method protects the privacy efficiently.	The proposed method can be implemented by taking different real

				product protocol and secure sum protocol		datasets.
10	Weiwei Fang, Bingru Yang, Dingli Song, Zhigang Tang	IEEE, 2009	Classification Method - Decision Tree	Homomorphic encryption technology	Results show that the new proposed algorithm provides good capability of privacy and security.	Proposed algorithm can be extended to other method of classification like clustering, association rule etc.
11	Sumana M, Hareesh K.S. and Shashidhara H.S.	ACM, 2010	Classification Method- Decision Tree	Scalar Product Protocol	Semi-trusted third party commodity server was used for privacy preservation.	As the number of parties increases, the communication cost also increases.
12	Vikas Ashok and Ravi Mukkamala	ACM, 2011	Association rule Mining	Deriving Association Rule locally and mining data with the help of third party	The scheme reduces the generation of spurious rules.	The filtering method used doesn't have theoretical proof which can be in further researches.
13	Gopal Behera	IEEE, 2011	Decision Tree Classifier	Secure multi-party computations for preserving privacy using C4.5 decision tree algorithm.	The proposed protocol of C4.5 gave better results as compared to ID3.	Alternate methods of SMC can be used for comparison of results.
14	Saeed Samet and Ali Miri	Elsevier, 2012	Neural network learning method	Back propagation algorithm and Extreme learning machine	The model is securely shared among all parties.	When volume of communication increases more efficient SVD algorithm can expand the proposed work.
15	Animesh Tripathy, Jayanti Dansana, Ranjita Mishra	ACM, 2012	Classification and Secure multi party computations	Contrast between gini index and entropy of attribute is presented.	Pruning of tree improves accuracy and privacy.	The tree constructed is very complex as it is very deep.
16	Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini	Springer, 2012	Anonymization	Game Theory and k-anonymity method	Used to find the consensual privacy protection level.	Comparison of the proposed method with other anonymization methods can be done for future work.
17	Jinfei Liu, Jun Luo, Joshua Zhexue Huang and Li Xiong	ACM 2012	Clustering Method	Density-based clustering algorithm-DBSCAN	Privacy of data is increased.	Further the proposed work can be extended by using other PPDM methods.
18	Hemlata Chahal	IJCA, 2013	Classification Method	Decision Tree	Without revealing the bank data, the algorithm predicts the credit risk of loan seekers.	The number of attributes can be increased and other methods of classification can be used for more efficiency.
19	Omar Abdel Wahab, Moulay	ACM,2014	Association Rule	Privacy-preserving of	Association rules queries are solved	The proposed technique can be further extended

	Omar Hachami et-al			distributed data using association rule mining approach-DRAM	efficiently and protects t inference attacks, preserves the privacy and confidentiality.	to other PPDM methods and results can be compared for efficiency.
20	Huafeng Ba, Xiaoming Gao, Xiaofeng Zhang and Zhenyu He	IEEE, 2014	Annonymization Method	k-annonymity and t-closeness approach by using IPFS and KIPFS algorithm for user privacy	Annonymizing the identified KIPFS, achieve better performance.	Further any existing privacy preserving algorithm can be integrated with the proposed algorithm for more efficiency and privacy.
21	Nasrin Irshad Hussain, Bharadwaj Choudhury and Sandip Rakshit	IJCA, 2014	Cryptographic technique	Encryption and key management was used for privacy. For clustering of raw data-set Rule-System was used.	New method of privacy preserving of Big data was proposed.	Efficiency and confidentiality can be increased by using different encryption methods for different types of data as compared to only one method (RSA method) in all situations.
22	H.R.Jalla and P.N. Girija	Springer, 2016	Walsh Hadamard Transformation(WHT) and perturbation technique	Combination of WHT and perturbation technique	The technique gave the results similar to K-Nearest Neighbour classifier	Approach applied only on horizontal partitioning of data sets.

PROPOSED ALGORITHM FOR DECISION TREE

Existing work presents a lot of technique for extracting and predicting knowledge form database while protecting sensitive data. Decision tree for PPDM can be proposed as an extended work. In this section an algorithm is proposed for creating decision tree of the vertically partitioned data.

A. ***Entropy and Information Gain***

There are many criterias for splitting of nodes among many given nodes. Here we consider the scheme given by Rastogi and Shim [24]. The entropy and Gini index are given as follows:

$$\text{Entropy}(S) = - p_p \log_2 p_p - p_n \log_2 p_n$$

Where p_p is the proportion of positive examples in S and p_n is the proportion of negative examples in S.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \varepsilon \text{Value}(A)} -|s_v|/|s| \text{ Entropy}(s_v)$$

B. ***Terms and Expressions used***

D1 represents first dataset owned by Owner1 and D2 represents the second dataset owned by Owner2.

A_r is an array of all the attributes of D1 and all the attributes of D2.

$A_r[m]$ is the m^{th} attribute in A_r.

N is a node which contains records of same class.

C. Algorithm

- Owner1 computes Information Gain of all the attributes owned by him i.e. D1 and Owner2 computes Information Gain of all the attributes owned by him i.e. D2.
- A semi- honest third party initializes the attribute with highest information gain as the root of the tree.
- Create a queue Q to contain the root.
- while Q is not empty do {
- Pop up the first node N from Q for each attribute $A_r[m]$ (for m=1…..k)
- Evaluate splits on attribute $A_r[m]$.
- Find the best split among $A_r[m]$'s.
- By using best split, split the node N into N1, N2…..Ns
- For i=1…..s, add N_i to Q if N_i is not well classified
- }

D. Flowchart

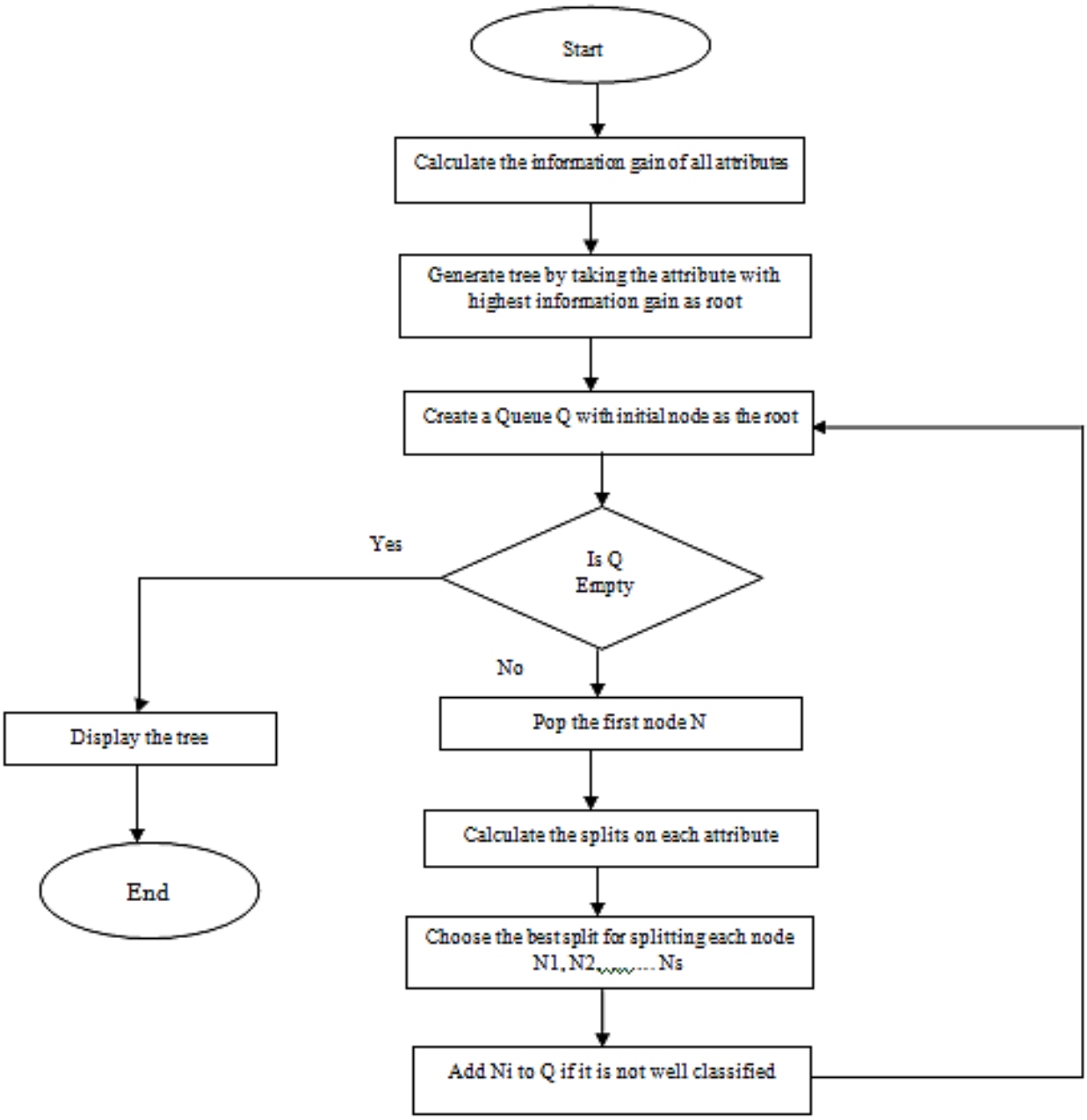


Figure 1. Flow chart of the proposed algorithm

CONCLUSION AND FUTURE WORK

By reviewing vast literature of PPDM, it was concluded that classification method can be used for protecting the private data from malicious users. So, decision tree method was used for predicting the class attribute. In the paper, a new algorithm is proposed to build a decision tree for vertically partitioned data. Implementation of the algorithm on real dataset can be done as an extended research work in future. The implementation can be done in R, SPSS, Matlab etc. for further research.

REFERENCES

[1] 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.

[2] Nan Zhang, Shenguan Wang and Wei Zhao, "A New Scheme on Privacy-Preserving Data Classification ∗ ", in the proceedings of KDD'05, August 21-24, ACM 2005.

[3] Animesh Tripathy, Jayanti Dansana, Ranjita Mishra, "A Classification Based Framework for Privacy Preserving Data Mining", in the proceedings of

ICACCI'12, August 3-5, ACM, 2012.

[4] Weiwei Fang, Bingru Yang, Dingli Song, Zhigang Tang, "A New Scheme on Privacy-preserving Distributed Decision-tree Mining", in the proceedings of First International Workshop on Education Technology and Computer Science, IEEE 2009.

[5] H.R.Jalla and P.N. Girija, "A Novel Approach for Horizontal Privacy Preserving Data Mining" , Advances in Intelligent Systems and Computing, pg 101-111, Springer 2016.

[6] Nasrin Irshad Hussain, Bharadwaj Choudhury and Sandip Rakshit, "A Novel Method for Preserving Privacy in Big-Data Mining", International Journal of Computer Applications(0975-8887) Volume 103-No 16, October 2014.

[7] Sumana M, Hareesh K.S. and Shashidhara H.S., "An Approach of Private Classification on Vertically Partitioned Data", in the proceedings of International Conference and Workshop on Emerging Trends in Technology(ICWET 2010), February 26-27, ACM 2010.

[8] Faris Alqadah and Raj Bhatnagar, "An Effective Algorithm for Mining 3-Clusters in Vertically Partitioned Data", in the proceedings of CIKM'08, October 26-30, ACM 2008.

[9] Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini, "Privacy Consensus in Anonymization Systems via Game Theory", Data and Applications Security and Privacy XXVI, proceedings of 26th Annual IFIP WG 11.3 Conference, DBSec 2012, published by Springer in July 2012.

[10] Omar Abdel Wahab, Moulay Omar Hachami et-al, "DARM: A Privacy-preserving Approach for Distributed Association Rules Mining on Horizontally-partitioned Data*", in the proceedings of IDEAS'14, July 07-09, ACM 2014.

[11] Vikas Ashok and Ravi Mukkamala, "Data Mining Without Data: A Novel Approach To Privacy-Preserving Collaborative Distributed Data Mining" in the proceedings of WPES'11, October 17, ACM 2011.

[12] Sheng Zhong and Zhiqlang Yang, "Guided perturbation: towards private and accurate mining" The VLDB Journal(2008) 17:1165-1177, Springer-Verlag 2007.

[13] Yi Xia, Yirong Yang and Yun Chi, "Mining Association Rules with Non-uniform Privacy Concerns" in the proceedings of DMKD'04 June 13, ACM 2004.

[14] Yanguang Shen, Hui Shao and Li Yang, "Privacy Preserving C4.5 Algorithm over Vertically Distributed Datasets" in the proceedings of "International Conference on Networks Security, Wireless Communications and Trusted Computing" pg. 446-448, IEEE 2009.

[15] Gopal Behera, "Privacy Preserving C4.5 Using Gini Index" 978-1-4244-9581-8, IEEE, 2011.

[16] *Emekci *, O.D. Sahin, D. Agrawal, A. El Abbadi, "Privacy preserving decision tree learning over multiple parties", Data & Knowledge Engineering 63 (2007) 348–361, Science Direct, Elsvier, 2007.*

[17] Weiwei Fang and Bingru yang, "Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data" in the proceedings of "International Conference on Computer Science and Software Engineering" IEEE, 2008.

[18] Saeed Samet and Ali Miri, "Privacy Preserving ID3 using Gini Index over Horizontally Partitioned Data" , 978-1-4244-1968-5/08 IEEE,2008.

[19] Saeed Samet and Ali Miri, "Privacy-preserving back-propagation and extreme learning machine algorithms", Data & Knowledge Engineering 79–80 (2012) 40–61, Elsevier,2012

[20] Huafeng Ba, Xiaoming Gao, Xiaofeng Zhang and Zhenyu He, "Protecting Data Privacy from being Inferred from High Dimensional Correlated Data" in the proceedings of "IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)", 978-1-4799-4143-8/14 IEEE 2014.

[21] Yanguang Shen, Hui Shao and Jianzhong Huang, "Research on Privacy Preserving Distributed C4. 5 Algorithm", in the proceedings of "2009 Third International Symposium on Intelligent Information Technology Application Workshops" 978-0-7695-3860-0/09 IEEE 2009.

[22] Hemlata Chahal, "ID3 Modification and Implementation in Data Mining", International Journal of computer Applications (0975-8887) Volume 80- No7, October 2013.

[23] Jinfei Liu, Jun Luo, Joshua Zhexue Huang and Li Xiong, "Privacy Preserving Distributed DBSCAN Clustering" in the proceeding of " PAIS 2012, March 30, ACM 2012.

[24] Rastogi, R. & Shim, K. (2000), '(public): A decision tree classifier that integrates building and pruning', Data Mining and Knowledge Discovery 4(4), 315–344.