Performance Evaluation of FP Growth algorithm in Privacy Preserving Data Mining using Hybrid Partitioning

Harpreet Kaur
Department of Computer Science and Engineering
DAVIET, Jalandhar, Punjab, India

Shaveta Angurala
Asst. Prof, Department of Computer Science and Engineering
DAVIET, Jalandhar, Punjab, India

Abstract- Data mining is the process of extracting useful information or knowledge from large data repositories and it is used in various business domains. In today's scenario there is a great need to protect the sensitive information from unauthorized access. In this paper, Hybrid technique for the distribution of data is used which is the combination of the horizontal and vertical data distribution. This technique is used for providing privacy to the data. FP Growth algorithm on hybrid partitioned dataset is used to provide accuracy and to decrease the execution time for generation of rules. The use of FP Growth algorithm reduces the utility loss of the data. The experiments are carried out on the two datasets namely adult and credit dataset and results are predicted on the basis of Apriori and FP Growth algorithm. The experimental results show that the FP Growth algorithm is better in performance than Apriori algorithm in terms of accuracy, execution time and utility loss.

Keywords – Accuracy, Apriori algorithm, Association rule mining, Execution Time, FP Growth algorithm, Hybrid Partitioning, Privacy preserving data mining, Utility Loss

I. INTRODUCTION

The process of extracting useful information or knowledge from large data repository is called data mining. In today's scenario the privacy of the data plays a vital role so that the third party does not access the sensitive information. Many techniques are applied to the data to provide privacy but still there is a problem in privacy preserving data mining that has been addressed by several researchers in the past [1], [2], [3], [4] there results are not able to provide privacy to data. Various data perturbation techniques are also used to provide privacy, but there results are also not fruitful.

Association rule mining is the major technique used in data mining for extracting frequent patterns and correlations from the transactional databases. [5], [6]. In association rule mining the relationships are identified between items. It uses two important basic measures. These are support and confidence [7]. These are the minimum conditions provided by the user.

FP Growth is an algorithm used for mining the frequent patterns. It uses depth-first search algorithm. FP Growth algorithm performs only two database scans. It does not generate any candidate set. For generating frequent patterns FP Growth algorithm uses divide and conquer method. [8], [19] and [20]. The performance of FP Growth algorithm is better than the apriori algorithm.

Horizontal and vertical partitioning are combined together to form the Hybrid partitioning. In vertical partitioning records are collected for a different set of attributes by the party, but each party have records for the same set of entities. In the horizontal partitioning records are collected for all of the attributes, but records are collected for the different entities [9]. By using vertical partitioning privacy to the data is provided and by using horizontal partitioning both the accuracy and privacy is achieved. Thus by using the hybrid partitioning whole data is protected from unauthorized access [16], [17] and [18].

The objective of this paper is to provide accuracy and privacy to the data. In this paper distributed technique is applied on the data. Horizontal and vertical partitioning is applied together, which form the hybrid partitioning. In the vertical partitioning, different subsets contain different attributes but have same records. In case of horizontal

partitioning, different subsets contain the same set of attributes, but have different records. Apriori algorithm is applied on vertical partitioned dataset and FP Growth algorithm is applied on both the vertical and hybrid partitioned dataset. Accuracy of the data is measured in terms of the proximity of the sanitized value to the original value. Utility loss is used to calculate how much information is lost during the sanitization of the data and also the execution time of the process is calculated. The comparison is done between the Apriori, FP Growth algorithm on vertically partitioned dataset and FP Growth on hybrid partitioned dataset on the bases of accuracy, utility loss and execution time.

The rest of the paper is organized as follows. : In section II related works is explained for defining the problem. Section III defines problem and work on that problem. In section IV new approach is proposed for generation of rules. Section V explained the experimental results. Section VI described the conclusion and section VII tells about the future scope of this work.

II. RELATED WORK

In the field of data mining the process of privacy preserving has played an important role. It is used for providing the security to the information so that the data remains protected from the unauthorized access without any loss of the information. In today's scenario people are aware for providing the security to the sensitive information or personal data, but still there is some loss of privacy has occurred which affects its security. Unintentional results are also generated due to the lack of privacy. Due to this reason several methods have been proposed for providing the privacy to the information or data. The results of privacy preserving data mining algorithms is explained in terms of its accuracy, data utility, performance, or level of uncertainty to data mining algorithms etc. There is no privacy preserving algorithms exists that exceed other algorithms on all possible criteria like utility, cost, complexity, performance, tolerance against data mining algorithms etc [14]. When the dataset is partitioned horizontally, security is not provided to the data in the distributed environment for privacy preserving association rule mining. Apriori and FP Growth algorithm is applied to the horizontally partitioned dataset for analyzing its performance and security. The results produced by both the algorithm shows that the FP Growth algorithm is better than the Apriori algorithm [11]. In case of vertical partitioning the accuracy of the data or information is lost. It affects the information produced after the partitioning of the data. To overcome this accuracy problem horizontal and vertical partitioning of the dataset is combined together to form the hybrid partitioning of the dataset. It provides accuracy in both the distributed and centralized scenario [12], [13]. Association rule mining is used to group the related items and preserving the individual data privacy without compromise the accuracy of global data mining task and global association patterns were driven from the distributed data. Global rules are generated after the vertical partitioning of the dataset and percentage of missed rules and percentage of spurious rules were calculated [10]. Associative classification is the technique used for mining class association rules. It has more advantages than the heuristic and greedy method because it removes noise easily and also it obtains higher accuracy. It generates the rule set which is more complete than the traditional classification method. The comparative analysis is performed on various associative classification techniques and also the comparison is performed on the basis of accuracy and efficiency by using the support and confidence as measures [21].When two party algorithm is used with minimum support level, it will efficiently discover frequent itemsets without revealing individual transaction values. It will achieve good individual security [15].

III. PROBLEM DEFINITION

In today's scenario, privacy preserving data mining has played a major role. Privacy is applied to the personal data so that the vast amount of data can be utilized for medical history, criminal records, shopping, credit and driving records. Some of the research areas which use the personal information are law enforcement, medical research and national security. For controlling the flow of the information privacy played the major role, in this way unauthorized people are not able to access the sensitive information.

There are various issues in privacy preserving data mining and we have to deal with all these issues. The major objective of privacy preserving data mining is to safeguard the sensitive information from the unauthorized access and also the utility of the data is preserved. In case of vertically partitioned privacy preserving data mining the rules are generated with the help of association rule mining on vertically partitioned dataset and then these rules are combined together to form the global rules and are passed to the data miner for the further processing of the data. In this process there is a loss of utility because during this process sensitive information is lost. Also in this process accuracy of the information is affected. As a result of this process unexpected predictions are made. During this process, use of apriori algorithm takes lot of time for generating the rules.

To overcome the limitations of accuracy, utility loss and execution time a new technique is proposed which involves hybrid partitioning of the dataset and the use of FP Growth algorithm. In case of apriori algorithm lot of information is lost in distributed environment. By using FP Growth algorithm no loss of information takes place i.e. the utility of

the data is preserved and the results produced are accurate and also the execution time is reduced for generating the rules.

II. Proposed Algorithm

The goal of the proposed algorithm is to provide privacy to the data in a distributed environment and maintain the accuracy of the data. Utility of the data is also preserved. Use of FP Growth algorithm helps in preserving the accuracy, utility loss and privacy of the data and generates results in small execution time.
Proposed approach is expressed with the help of following steps:
Step1: Collect datasets from the UCI machine learning repository. In this algorithm two datasets are collected namely Adult and Credit dataset and convert them into the binary format by applying the filters. After its binary conversion in case of adult dataset 14253 attributes are generated but consider only 25 attributes, in case of credit dataset 1075 attributes are generated but we consider only 38 attributes.
Step 2: Apply preprocessing and filtering process on the dataset. In the filtering process filters are applied to the dataset so that it will convert into the binary form. Apply NominalToBinary and NumericToBinary filters. There filters are applied by using the weka tool.
In the preprocessing, values of minimum support and minimum confidence are provided. It is calculated as:

$$support(A \Rightarrow B) = \frac{\#tuples\ containing\ both\ A\ and\ B}{total\ \#\ of\ tuples}$$

$$confidence(A \Rightarrow B) = \frac{\#tuples\ containing\ both\ A\ and\ B}{\#tuples\ containing\ A}$$

Step 3: Apply FP Growth algorithm for the generation of the rules. Once the rules are generated these are stored for the further processing.
Step 4: Make two subsets from the original dataset by dividing the original dataset horizontally and apply FP-Growth algorithm on reach partition for the generation of rules. Figure 1 shows the horizontal partitioning of the dataset. In this approach two horizontal partitions are made

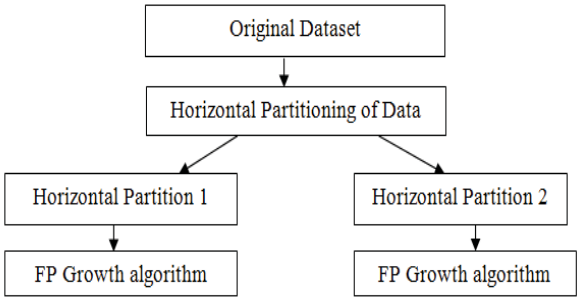

Figure 1: Horizontal Partitioning

Step 5: Apply vertical partitioning on each horizontal subset and make three vertical subsets after the partitioning FP Growth algorithm is applied on each vertical subset for the generation of rules and also calculate the execution time of each partition individually. Figure 2 shows the vertical partitioning of the dataset.

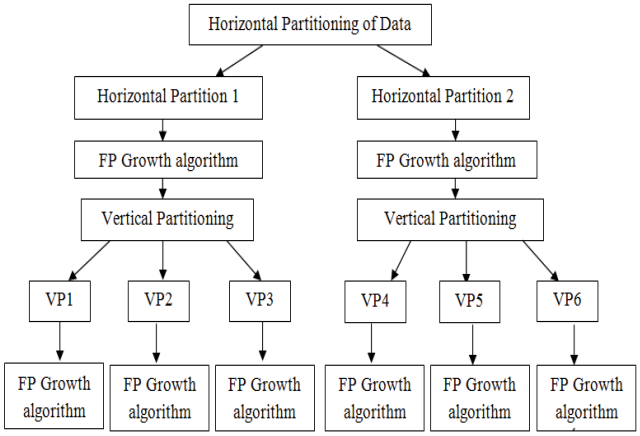

Figure 2: Vertical Partitioning

VP1 is the vertical partition1, VP2 is the vertical partition2, VP3 is the vertical partition3, VP4 is the vertical partition4, VP5 is the vertical partition5, and VP6 is the vertical partition6.

Step 6: Combine all the horizontal and vertical partitions together to generate global rules and also combine the execution time of each partition to generate total execution time. Global rules are provided to the third party

Global_Rules= Local_Rules+(i+1)+ " . " + tokens[i] +

Total_Time= t1+t2+t3+t4+t5+t6

Step 8: Evaluate the accuracy and the utility of the data. FP Growth algorithm on hybrid partitioned dataset shows better accuracy in comparison with the FP Growth and Apriori algorithm on vertically partitioned dataset. Also the utility loss of FP Growth algorithm on Hybrid partitioned dataset is less.

$$accuracy = \frac{n}{l} * 100$$

$$utility\ loss = \frac{u}{l} * 100$$

$$n = l - m$$

$$u = m - o$$

Where n is the not matched rules, l is the total rules, u is the utility, m is the matched rules and o is the original rules.

V. EXPERIMENTAL RESULTS

For evaluating the accuracy, utility loss and execution time, the experiments are carried out on two datasets namely adult and credit dataset which are taken from UCI Machine learning repository. For performing the experiment, netbeans tool is used for performing this experiment. First, collect the datasets and convert them into the binary form by using the weka tool. In the adult dataset 25 attributes and 126 instances are considered and in credit dataset 38 attributes and 1000 instances are considered.

For this experiment hybrid data distribution is used. First, two horizontal subsets are made from the original dataset and then each horizontal subset has three vertical partitioned subsets. After the partitioning, rules are combined together to form global rules. Then calculate accuracy, utility loss and execution time. Thus, FP Growth shows better accuracy and has small utility loss and also it has small execution time.

In case of adult dataset accuracy utility loss and execution time corresponding to the different support value is explained in Table 1, Table 2 and Table3. The dataset is tested using two different algorithms, these are Apriori and FP Growth, and two different partitions that is vertical and hybrid partition and. Figure 3, Figure 4 and Figure 5 shows the comparison of accuracy, utility loss and execution time graphically.

Table 1. Accuracy for Adult dataset

Support	Accuracy in %age		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	35.48	91.6	98.73
0.2	33.36	64.52	97.43
0.3	33.36	60.71	94.53

Table 2. Utility Loss for Adult dataset

Support	Utility Loss in %age		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	53.76	4.58	0.55
0.2	66.67	19.35	1.1
0.3	33.36	21.43	2.34

Table 3. Execution Time for Adult dataset

Support	Execution Time (ms)		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	2240	15	16
0.2	2240	16	15
0.3	2355	31	16

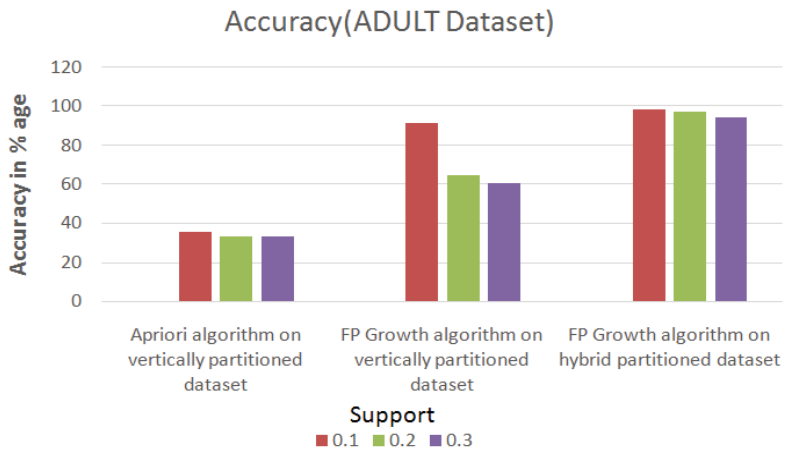

Figure 3: Comparison of Accuracy on Adult dataset

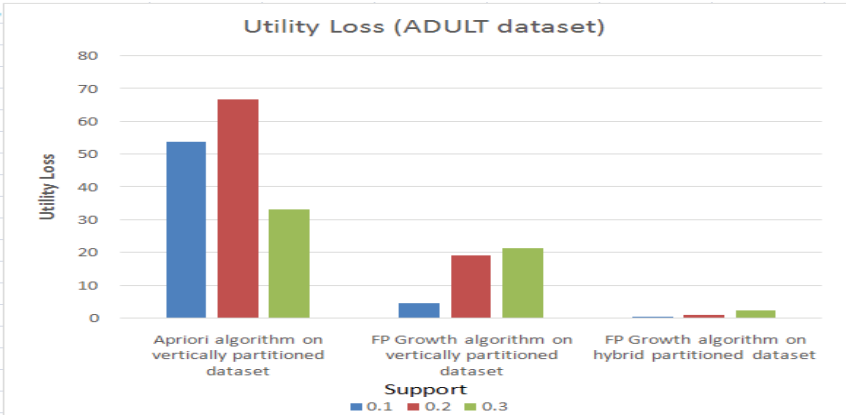

Figure 4: Comparison of Utility Loss on Adult dataset

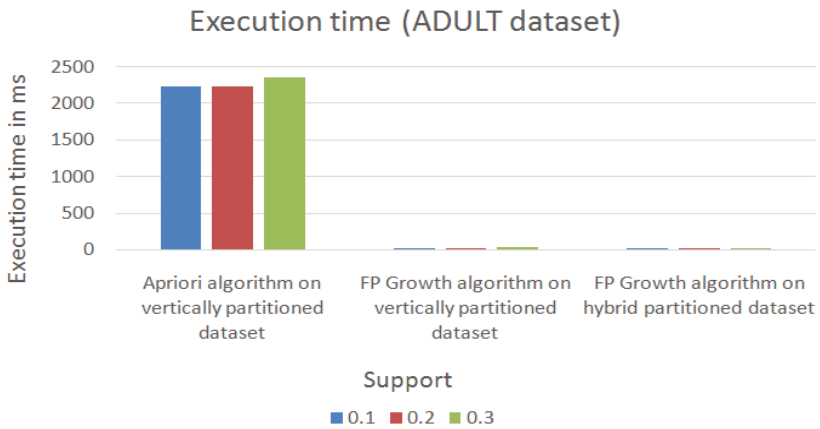

Figure 5: Comparison of Execution Time on Adult dataset

Credit dataset is expressed in Table 4, Table 5 and Table 6 in terms of its accuracy, utility loss and execution time. Figure 6, Figure 7 and Figure 8 graphically shows the comparison of accuracy, utility loss and execution time.

Table 4. Accuracy for Credit dataset

Support	Accuracy in %age		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	33.36	97.95	99.33
0.2	33.36	77.58	97.39
0.3	33.36	50	94.93

Table 5. Utility Loss for Credit dataset

Support	Utility Loss in %		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	33.36	1.26	0.28
0.2	66.67	13.79	1.12
0.3	33.36	30.77	2.17

Table 6. Execution Time for Credit dataset

Support	Execution Time (ms)		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	1124	48	47
0.2	1170	47	16
0.3	1155	48	45

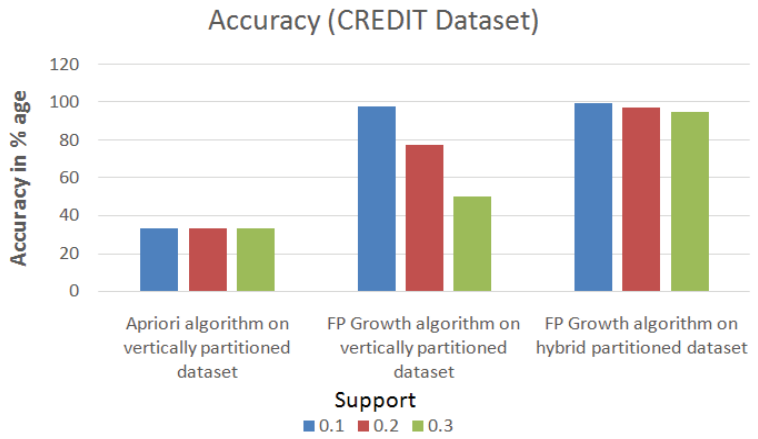

Figure 6: Comparison of Accuracy on Adult dataset

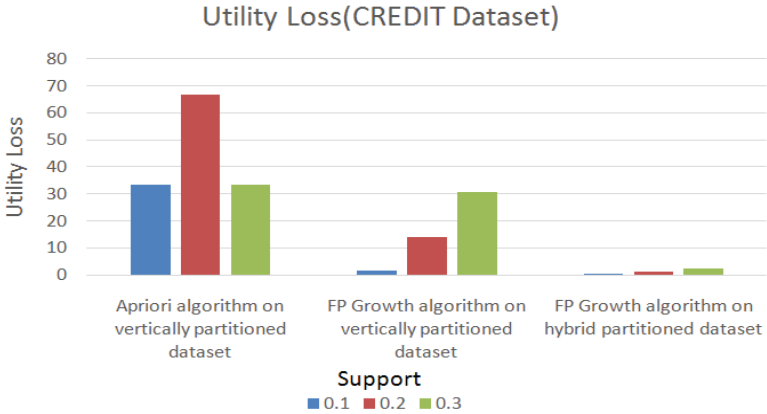

Figure 7: Comparison of Utility Loss on Credit dataset

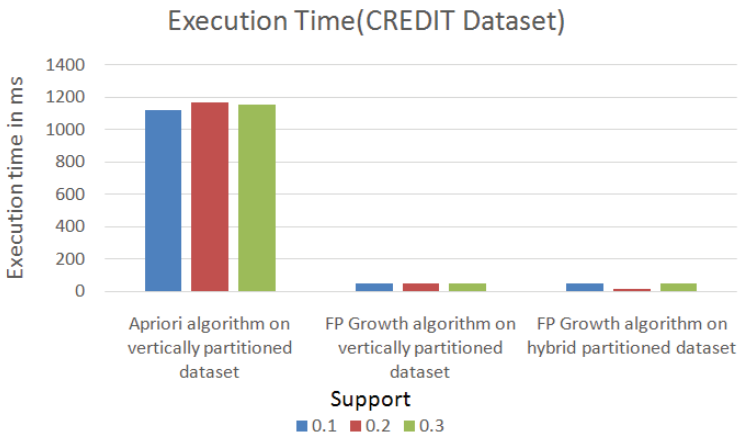

Figure 8: Comparison of Execution Time on Credit dataset

VI. CONCLUSION

The proposed work shows that use of FP Growth algorithm on hybrid partitioned dataset can provide accuracy to the data also have less utility loss that is no useful information is lost, and provide rules in small execution time. In FP Growth algorithm useful information is preserved as compared to the Apriori algorithm. In FP Growth algorithm candidate set is not generated while in Apriori algorithm candidate set is generated due to this reason it will take large execution time. FP Growth uses divide and conquer method for generating frequent items. Accuracy, utility loss and execution time is observed efficiently by considering the different values of support. The comparison of

accuracy, utility loss and execution time is done between the Apriori algorithm on vertical partitioned dataset, FP Growth algorithm on vertical partitioned dataset and FP Growth algorithm on hybrid partitioned dataset. Results show that as the value of minimum support is increased, accuracy of the data decreases and utility loss increases.

VII. FUTURE SCOPE

In future we would like to work on hybrid partitioned dataset and to improve FP Growth algorithm to enhance its performance.

REFERENCES

[1] K. Liu, H. Kargupta and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowledge and Data Engg, 18(1):92-106, January 2006.

[2] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), pages 24-31, June 2 2002.

[3] Benjamin C. M. Fung , Ke Wang , Rui Chen , Philip S. Yu, Privacy preserving data publishing: A survey of recent developments, ACM Computing Surveys (CSUR), vol. 42, no. 4, pp. 1-53, 2010.

[4] Yin, Yong, Ikou Kaku, Jiafu Tang, and JianMing Zhu. "Privacy preserving Data Mining," In Data Mining, pp. 101-119. Springer London, 2011.

[5] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," In Proc. Conf. Management Data ACM SIGMOD, pp. 1–12, 1996

[6] K.Saranya, K.Premalatha and S.S.Rajasekar, "A Survey on Privacy Preserving Data Mining" , IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEM (ICECS 2015)

[7] T. Karthikeyan and N. Ravikumar, "A Survey on Association Rule Mining," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), pp. 5223-5227, 2014.

[8] Anusuya M ,Sudharani K ,Ganthimathi M ,Sumathi G, "Frequent Itemset Mining Using PFP-Growth via Transaction Splitting", International Journal of Innovative Research in Computer and Communication Engineering, An ISO 3297: 2007 Certified Organization, Vol. 4, Issue 2, February 2016

[9] Jaideep Vaidya, Senior Member, IEEE, Basit Shafiq, Member, IEEE, Wei Fan, Member, IEEE, Danish Mehmood, and David Lorenzi, "A Random Decision Tree Framework for Privacy Preserving Data Mining", IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2014

[10] Vikas G. Ashok, K. Navuluri A. Alhafdhi R. Mukkamala, "Dataless Data Mining: Association Rules-based Distributed Privacy-preserving Data Mining", 2015 12th International Conference on Information Technology - New Generations

[11] Patil Suraj K, Gadage Shrinivas, "Privacy Preserving Two Party Distributed Association Rule Mining by FP Growth on Horizontally Partitioned Data", International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 3, Issue 6, June 2015

[12] Asha Khatri, Swati Kabra, Shamsher Singh and Durgesh Kumar Mishra, "Architecture for Preserving Privacy During Data Mining by Hybridization of Partitioning on Medical Data", 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation

[13] M. Saravanan, A. M. Thoufeeq, S. Akshaya & V.L. Jayasre Manchari, "Exploring New Privacy Approaches in a Scalable Classification Framework", Data Science and Advanced Analytics (DSAA), 2014 International Conference

[14] Majid Bashir Malik , M. Asger Ghazi and Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", 2012 Third International Conference on Computer and Communication Technology

[15] Jaideep Vaidya, Chris Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining in 2002

[16] Chris Clifton, "Privacy preserving distributed data mining" In ACM SIGKDD Explorations, November 9, 2001

[17] Gang Kou, Yi Peng1, Yong Shi2, and Zhengxin Chen, "Data mining of medical data using data separation-based technique" Data Science Journal, volume 6, supplement, 30 July 2007, pp S429-S434.

[18] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis "State-of-the-art in Privacy Preserving Data Mining" In the proceeding of SIGMOD Record, Vol. 33, No. 1, March 2004, pp 50-57

[19] DANIEL HUNYADI, "Performance comparison of Apriori and FP-Growth algorithms in generating association rules", Proceedings of the European Computing Conference, Department of Computer Science"Lucian Blaga" University of Sibiu, Romania

[20] Abdullah Saad Almalaise Alghamdi, "Efficient Implementation of FP Growth Algorithm-Data Mining on Medical Data", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.12, December 2011

[21] Nitendra Kumar Vishwakarma, Jitendra Agarwal, Shikha Agarwal, Sanjeev Sharma , "Comparative Analysis of Different Techniques in Classification Based on Association Rules", 2013 IEEE International Conference on Computational Intelligence and Computing Research.