Privacy Preserving in Data Mining using FP Growth Algorithm on Hybrid Partitioned Dataset

Harpreet Kaur
M.Tech, CSE
DAVIET, Jalandhar

Shaveta Angurala
Assistant Prof. CSE
DAVIET, Jalandhar

ABSTRACT
Data mining is used in various business domains to extract important information from the large data repositories. In this paper, Horizontal and Vertical data distribution is combined to provide privacy to the data. FP Growth algorithm on hybrid partitioned dataset is used to decrease the execution time for generation of rules. 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 execution time because FP Growth algorithm takes less time to generate rules.

Keywords
Apriori algorithm, Association rule mining, FP Growth algorithm, Hybrid Partitioning, Privacy preserving data mining

1. INTRODUCTION
Data mining is defined as the process of extracting useful knowledge or information from large data repository. To provide privacy to the data is the major issue so that the third party is not able to access the sensitive information. There is a problem in privacy preserving data mining that has been addressed in the past by the several researchers [1], [2], [3], [4] but the results are not able to provide privacy to the data. Data perturbation techniques are also used in the past to provide privacy but they are also not providing useful results.

FP Growth algorithm is used for mining the frequent patterns which use depth-first search algorithm. FP Growth algorithm do not generate any candidate set. It performs only two database scans which make it faster than the Apriori algorithm. For generating frequent patterns FP Growth algorithm uses divide and conquer method. [5], [16] and [17]

Hybrid partitioning is the combination of horizontal and vertical partitioning. In the horizontal partitioning, parties have records for all of the attributes but collect records for different entities. In case of vertical partitioning, party's collects records for a different set of attributes but each party have records for the same set of entities [6]. In case of vertical partitioning privacy is achieved and in horizontal partitioning both privacy and accuracy is achieved [13], [14], and [15]

The objective of this paper is to apply distribution technique on data for preserving the privacy of the data. Horizontal and Vertical partitioning of the dataset are combined together to form Hybrid partitioning. In Horizontal Partitioning different subsets contains the same set of attributes but having the different records while in Vertical Partitioned dataset different subsets contains the different attributes but having the same records. On vertical data distribution both Apriori and FP Growth algorithms are applied. FP Growth algorithm is applied on Hybrid partitioned dataset and compare the results of Apriori algorithm, FP Growth algorithm for vertical partition with FP Growth for hybrid partitioned datasets on the basis of execution time.

The flow of the paper is as follows: In section 2 related works is explained for defining the problem. In section 3 problem is defined and work on that problem. In section 4 new approach is proposed for generation of rules. Section 5 explained the experimental results. Section 6 described the conclusion and section 7 tells about the future scope of this work.

2. RELATED WORK
In data mining the process of privacy preserving has played a vital role. It helps in providing the security to the sensitive information or knowledge and protecting information from unauthorized access without affecting the security of the data. Now a day's people are aware of the privacy intrusions on their personal data and they do not share their sensitive information to unauthorized people. Lack of privacy may generate the unintentional results. Several methods have been proposed in privacy but still it has its significance. The results of privacy preserving data mining algorithms is explained in terms of its 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 [11]. In case of horizontally partitioned dataset the security is not provided for distributed privacy preserving association rule mining. Apriori and FP Growth algorithm are applied to analyze the performance and security. The results produced by the FP Growth algorithm are better than the Apriori algorithm [8]. The combination of the horizontal and vertical partitioning of the dataset is known as the hybrid partitioning. When privacy is provided to both horizontal and vertical partitioned dataset in distributed and centralized scenario can improve the accuracy which overcomes the accuracy problem in the vertical partitioning [9], [10]. 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 [7]. 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 [12].

3. PROBLEM DEFINITION
Today, the major concern is about the privacy of the personal data so that vast amount of the data can be utilized efficiently such as shopping habits, credit and medical history, criminal records, and driving records according to requirements. This personal information is used in many research areas such as medical research, law enforcement and national security. By providing the privacy to the data flow of the information can be controlled.

There are various privacy issues in the data mining and there is a need to deal with all these issues. In recent years, a subfield of data mining called privacy preserving data mining has gained a great development. The objective of privacy preserving data mining is to protect the sensitive information from unauthorized or unsanctioned access and preserve the utility of the data. The process of privacy preserving data mining is two-fold. First, the spurious rules are generated with the application of Association rule mining on Subset of Vertically partitioned data. Secondly these rules are combined and forwarded to the data miner where actual process of data mining takes place. Although this technique helps to maintain the privacy of the individual's data during the process of data mining, there are certain limitations associated with it. The most prominent limitation is the loss of utility, which result in unexpected predictions from data mining due to the perturbation of data. The other limitation is the execution time taken by the association algorithm to generate rules of the larger dataset.

To overcome these limitations a new technique is proposed which involves hybrid partitioning of the dataset. The use of FP-Growth algorithm instead of Apriori, results in reduction of execution time taken to generate rules of larger datasets.

4. PROPOSED APPROACH

The goal of the proposed algorithm is to generate global rules from the hybrid partitioned data in small execution time and also maintain the privacy of the data.

The proposed approach is as follows: The data is collected from the repository by the data owner. In this approach the experiment is carried out on two datasets namely adult and credit dataset. Apply preprocessing and filtering process by providing the different values of minimum support and minimum confidence. Consider the values of minimum support as 0.1, 0.2, 0.3, 0.4, and 0.5. So that rules are generated efficiently. The filtering process validates the support. FP Growth algorithm is used for the generation of the rules. Once the rules are generated these are stored for the further processing by the data owner.

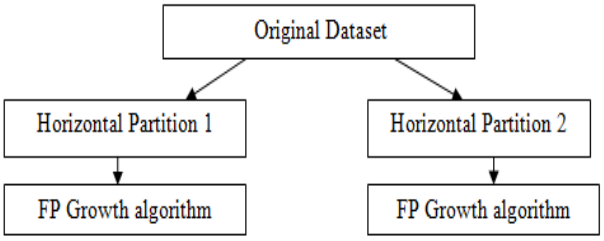

Figure 1: Horizontal Partition

Now make two subsets from the original dataset by dividing the original dataset horizontally. Figure 1 shows the horizontal partitioning of the dataset. Apply FP-Growth algorithm on reach partition for the generation of rules.

After horizontal partitioning apply vertical partitioning on each horizontal subset and make three vertical subsets. Fig 2 shows the vertical partitioning of the dataset. Apply FP Growth algorithm on each vertical partition and also calculate the execution time of each partition individually. In Figure 2 VP1 is the vertical partition1, VP2 is the vertical partition2 and VP3 is the vertical partition3. This partitioning is done on Horizontal partition1. Similarly, the partition is done on Horizontal Partition 2 and rules are generated by applying FP Growth algorithm.

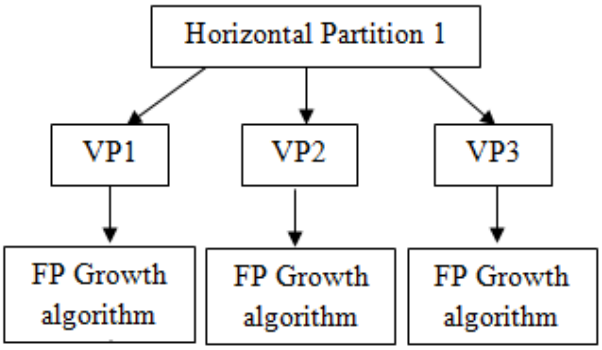

Figure 2: Vertical Partition

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. Evaluate the parameters such as percentage of missed rules and percentage of spurious rules.

5. EXPERIMENTAL RESULTS

To evaluate the execution time the experiments are carried out on two datasets namely adult and credit dataset which are taken from UCI machine learning repository. First, convert the dataset in binary form by applying the filtering process and then the datasets are used. In case of adult dataset 25 attributes and 126 instances are considered. In case of credit dataset 38 attributes and 1000 instances are considered. Different support values are used for the calculation of the execution time.

For the preservation of the privacy the hybrid distribution that is the combination of the horizontal and vertical partitioning is done so that the third party is not able to recognize the original data. The original dataset is partitioned horizontally to make horizontal subsets. It has different records. After the horizontal partitioning each horizontal subset is further partitioned vertically to make vertical subsets. In this case two horizontal subsets are made and each horizontal subset has three vertical subsets. After the partitioning of the dataset the execution time is calculated individually and then the execution time are combined to calculate the overall execution time and also the rules are combined so that these rules are provided to the third party.

Adult dataset: In case of adult dataset execution time corresponding to the different support value is explained in Table 1. The dataset is tested using two different partitions that is vertical and hybrid partition and also two different algorithms that is Apriori and FP Growth algorithm. Figure 3 shows the comparison of execution time graphically.

Table 1. Execution Time for Adult dataset

Support	Execution Time (ms)		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	217	40	18
0.2	150	17	10
0.3	135	15	3
0.4	135	14	5
0.5	121	14	2

Credit Dataset: Credit dataset is expressed in Table 2 in terms of its execution time. Figure 4 graphically shows the comparison of execution time.

Table 2. Execution Time for Credit dataset

Support	Execution Time (ms)		
	Apriori on Vertical	FP Growth on Vertical	FP Growth on Hybrid
0.1	321	107	79
0.2	266	102	22
0.3	280	87	21
0.4	319	78	22
0.5	313	66	19

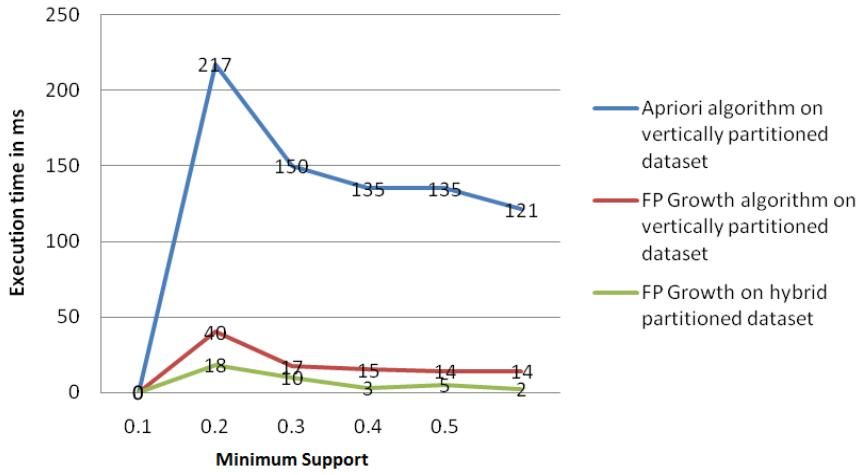

Figure 1: Comparison of Execution Time on Adult dataset

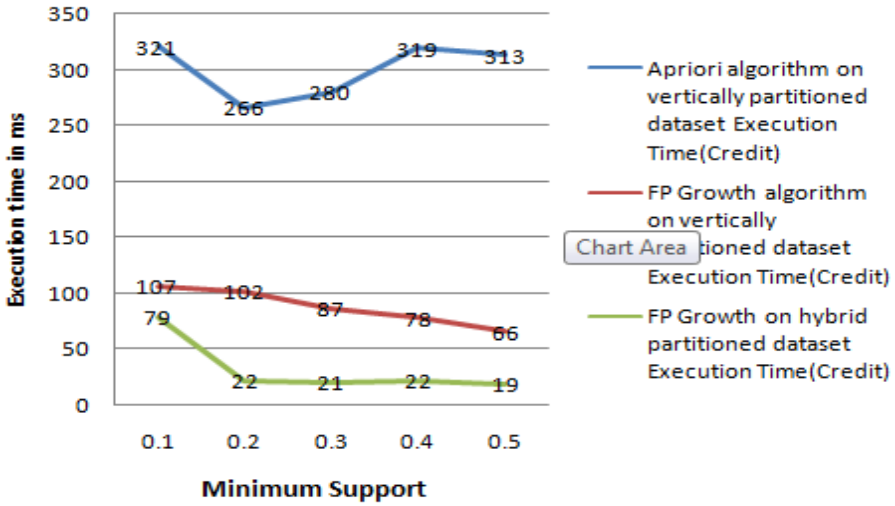

Figure 2: Comparison of Execution Time on Credit dataset

6. CONCLUSION

The proposed work shows that the use of FP Growth algorithm on hybrid partitioned dataset can decrease the execution time. In FP Growth algorithm candidate set is not generated. In case of Apriori algorithm candidate set is generated which results in large execution time. FP Growth uses divide and conquer method for generating frequent items. Execution time is observed efficiently by considering the different values of minimum support. The comparison of 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 the execution time is decreased. Privacy of the data is preserved with the combination of horizontal and vertical partitioning.

7. FUTURE SCOPE

In future we would like to work on the Hybrid partitioned dataset by applying FP Growth algorithm so that the accuracy of the data is preserved and the utility loss of the data is decreased so that quality of the data is preserved.

8. ACKNOWLEDGMENTS

I am grateful to Ms. Shaveta Angurala, Assistant Professor at DAVIET, Jalandhar. Under her complete guidance and support I became able to complete my research paper.

9. 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] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

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

[14] 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*.

[15] 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.

[16] 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.

[17] 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