Efficient Encrypted Data Distribution Vertically by Generating Frequent Pattern

Preeti Pal Singh
(Mtech Scholar CSE Dept.)
Sagar Institute Of ResearchTechnology and Excellence,RGPV University

Anjana Verma
(Prof. CSE Dept.)
Sagar Institute Of ResearchTechnology and Excellence,RGPV University.

ABSTRACT
With the rapid increase in digital world servers security for data is highly required. So most of researcher have proposed different techniques such as data partition and modification. Here proposed work has resolve this issue of digital data security by vertical partition and AES encryption algorithm. In this work vertical patterns are generate from the database by the use of aprior algorithm of association rule mining. These patterns effectively distribute data for different sites. While ach site maintain an index table of inserted rows for proper database operations. Experiment is done on real adult dataset. Results shows that proposed work is better as compare to previous existing algorithm on different evaluation parameters.

Keywords
Conditional Functional Dependency, Data Anonymization, Encryption, Effective Pruning,

1. INTRODUCTION
In Mining knowledge discovery plays an important role for investigating raw data. Different researchers are working on this where new patterns and unknown information can be extract from large set of data. So data collection, dataset elimination are new field of research for maintain the data privacy with consistency. Here this is required because current data mining algorithms are so intelligent that information can be easily removed, learn or modify. So this kind of study or work comes under privacy preserving mining. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking. Privacy preserving data mining [9, 8], is a novel research direction in data some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem. In the similar fashion, this work provide a new data partition technique where perfect partition is done with high rate of accuracy for privacy on various servers. The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information.

As various approaches has been found in this field of privacy preserving mining in last few years [5, 9]. Furthermore, the difficulty has been discussed in numerous communities such as the cryptography community, the numerical revelation control community and the database community. Data mining techniques have been developed effectively to extracts information in order to sustain a variety of domains advertising, weather predictions, medical analysis, and national safety. But it is still a challenge to mine certain kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing problem in health care applications.

Privacy-preserving data mining has emerged to address this issue, with several papers in the past few years as well as articles in the popular press[2,3]. One approach is to alter the data before delivering it to the data miner. The second approach assumes the data is distributed between two or more sites, and these sites cooperate to learn the global data mining results without revealing the data at their individual sites. This approach was first introduced to the data mining community, with a method that enabled two parties to build a decision tree without either party learning anything about the other party's data, except what might be revealed through the final decision tree. This work have since developed techniques for association rules[2, 4], decision trees with vertically partitioned data, clustering, k-nearest neighbor classification, and other methods are in progress.

There are hundreds of works proposed for providing privacy for the data for different requirements. So classification of those methods are done on the basis of following measures:

- Data sharing
- Data Perturbation
- Data Mining Algorithm
- Rule suppressing
- Privacy conservation.

The first point discuss the sharing of data. Some of the approach have been proposed and developed for data centralization, while others refer to a data sharing situation. Data Sharing situation can further be divide into horizontal data partition and vertical data partition. Horizontal partition refers to these cases where data is divide row wise on different servers or stations or datacenters. While vertical data partition, refers to the cases where all the data is divide in column wise on different servers or stations or datacenters. The second dimension discusses to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection [6, 8]. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization.

2. RELATED WORK
Chiang & Miller (2008) have proposed an algorithm that find the different conditional functional dependencies of the dataset. So those rules that harm the identity of the individual is filter out. So this reduce the distance between the dependencies and association rules. So this study help in

identifying the semantic rule for broad range of constraints to apply.

Li et al (2013), problem of finding the minimal set of constants for conditional functional dependency present in used dataset. Here minimal set of conditional functional dependency is obtained by minimal generator as well as by clousers of those sets. Here proposed work has find the pruning criteria so overall work get reduce and unwanted generator, closures get shorten. So based on the proposed work a dataset modal is generate where each node act as a data row. Pruning of node is depending on two condition first is node have no conditional functional dependency rules. Second is descendent node of the node have no conditional functional dependency rules.

Yka Huhtala et. al. in [5], has proposed a work that generate conditional functional dependency and approximation rules by utilization of partitions. So by dividing the large dataset in to some partitions generation or searching of functional rules get easy and accurate.

Hong Yao [7] has developed an algorithm named as FDMine (Functional Dependency Mining). Here FD-Mine develops rules by utilizing the functional dependency properties in theory which reduce dataset size for searching as well as filter some of the unwanted or unfruitful rules. It has also proved in the work that pruning of rules not lead to loss of information in the work. Here whole work experiment is done on IS UCI datasets. Here pruning of rules are more as compare to previous works while evaluating results get improved.

3. PROPOSED WORK

Whole work is divide in two module first is site architecture building then data distribution as per proposed architecture. Here as per the relation between the data items distribution of data is done. Before transferring the data to the site encryption is performed. Explanation of whole work is shown in fig. 1.

3.1 Pre Processing

Pre-Processing: As the dataset is obtain from the above steps contain many unnecessary information which one need to be removed for making proper operation. Here data need to be read as per the algorithm such as the arrangement of the data in form of matrix is required.

3.2 Generate Rules

In order to hide the information from the dataset one approach is to reduce the support and confidence of the desired item. For finding the item set which is most desired one has to find that the frequent pattern in the dataset. There are many approaches of pattern finding in the dataset which are are most frequent one of the most popular is aprior algorithm is use in this work. In aprior algorithm all possible rules are generate from the dataset by combining different values of the one coloumn to the other. Such as A1→B1, A1→B2, A1→D1,……A1→E1, A1→E2, A1→E5. In similar fashion other rules are prepared.

Table 1. Assumed Data for distribution

C1	C2	C3	C4
A1	B1	D1	E1
A2	B2	D2	E1
A3	B3	D3	E1
A4	B3	D4	E4
A5	B4	D5	E5

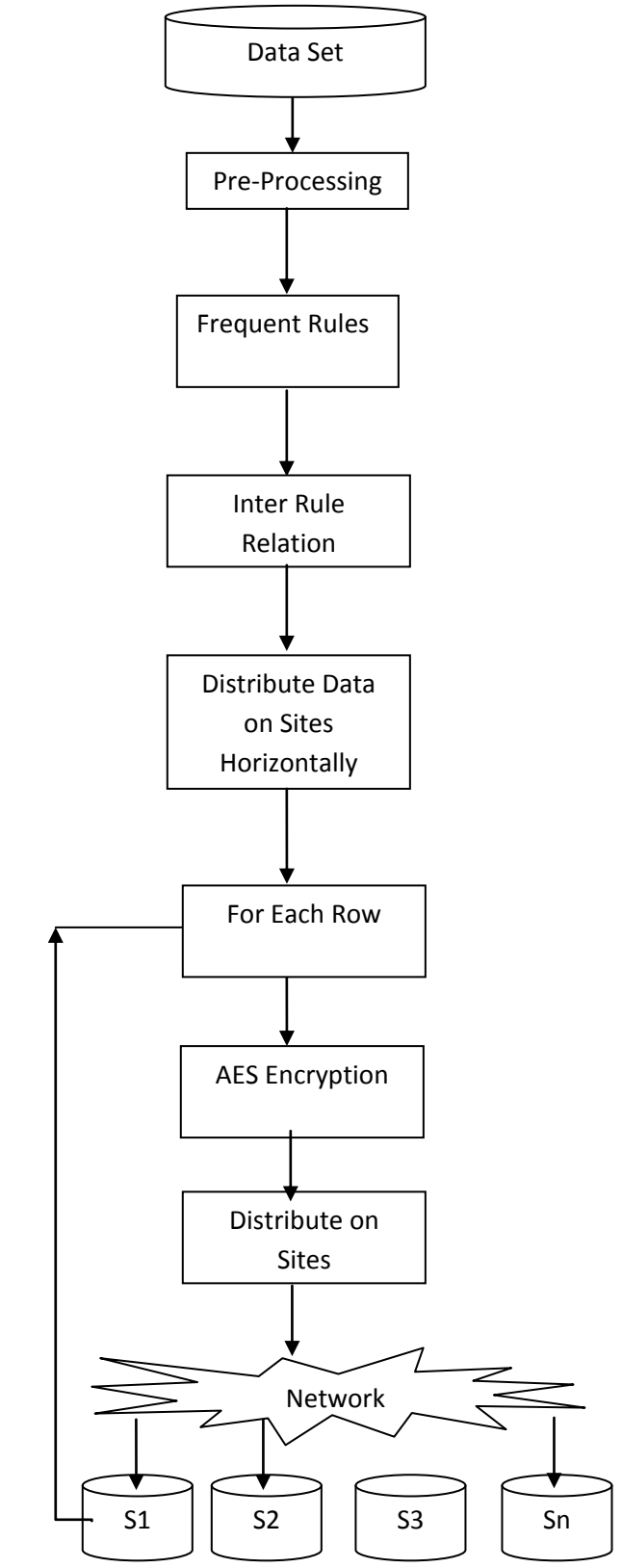

Fig. 1 Proposed Work Block Diagram.

3.3 Inter Rule Relation

Now as per the different frequent rules of the dataset the relation between columns can be evaluate. Here all possible

pair of columns are prepared then find number of rules between them such as CP={(C1,C2), (C1,C3), (C1,C4), (C1,C5), (C2,C3), (C2,C4)……(C4,C5)}. This can be understand as column pattern (C1,C2) have $\{I_i \rightarrow I_j\}$ where i=(A1, A2, A3} and j=(B1, B2, B3}. So if total rules present in the dataset column act as the bond strength between the columns. Sort Highly related Rules in other words pattern having highest number of rules in there group of columns is consider as the strongly related column group.

3.4 Distribute Data On Site

Here in this step whole columns as per there bonding with other column is distribute on the different site. Here it was try to put strongly column on single site but due to limitations of the site storage, low bonded column is distribute to other site. So depend upon the relation between the columns data partition is done.

Table2. Data Column partition on different sites

Site	Columns		
S1	C3	C5	C1
S2	C6	C2	
S3	C3		

3.5 Insertion of Rows

For each row in the data before transferring data to the selected site it need to be encrypt first. Here **AES** (Advanced Encryption Algorithm is use. In this algorithm 128 bit key is use with 16 character as a plaintext. Total 9 rounds are apply on the plaintext matrix having four common steps first is substitution, second is row shift, third is poly-matrix XOR operation, while fourth is round key XOR operation.

4. EXPERIMENT AND RESULT

This section presents the experimental evaluation of the proposed perturbation and de-perturbation technique for privacy prevention. To obtain AR this work used the Apriori algorithm [1], which is a common algorithm to extract frequent rules. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

4.1 Data Set

In [9] it has use Adult dataset where it contain different discriminating item set such as country, Gender, Race, 1996. This data set consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The data set has 14 attributes (without class attribute). For this work experiments with the Adult data set is use.

4.2 Evaluation Parameters

4.2.1 Elapsed Time

Here total execution time (second) is calculate for the data distribution on different sites. Two type of elapsed time is calculate over here first is Incremental Time and other is Batch time.

Incremental Time: Time required for distribution of single row data on different site is termed as Incremental Time.

Batch Time: Time required for distribution of Batch of data on different site is termed as Batch Time.

4.2.2 Space Cost

As data is distributed as per the pattern in the dataset so a perfect pattern have less number of combinations to represent same data. So number of cells required for the storage of data on different sites is termed as Space Cost.

4.3 Result

Table 3. Comparison of Average Incremental Elapsed time

Dataset Size	Proposed Work	Previous Work
300	**0.003859**	**0.004796**
400	**0.004052**	**0.004362**
500	**0.004262**	**0.004138**

From above table 3 it is obtained that proposed work is better as compare to previous work in [13]. As average incremental elapsed time is less while executing proposed work algorithm. It has seen that by increase in dataset size is remain almost same.

Table 4. Comparison of Average Batch Elapsed time.

Dataset Size	Proposed Work	Previous Work
300	**0.012974**	**0.01807**
400	**0.017063**	**0.02050**
500	**0.02186**	**0.02325**

From above table 4 it is obtained that proposed work is better as compare to previous work in [13]. As average batch elapsed time is less while executing proposed work algorithm. It has seen that by increase in dataset size also increases. Time requirement is less because of generations of perfect patterns in proposed work as this reduce pattern size.

Table 5. Comparison of Average Space Cost.

Dataset Size	Proposed Work	Previous Work
300	**1728**	**1808**
400	**2168**	**2272**
500	**2588**	**2740**

From above table 5 it is obtained that proposed work is better as compare to previous work in [13]. As average space required for proposed work algorithm is comparatively less. Space requirement is less because of generations of perfect patterns in proposed work. It has seen that by increase in dataset size is space cost also increases but not in same ratio as data set increases.

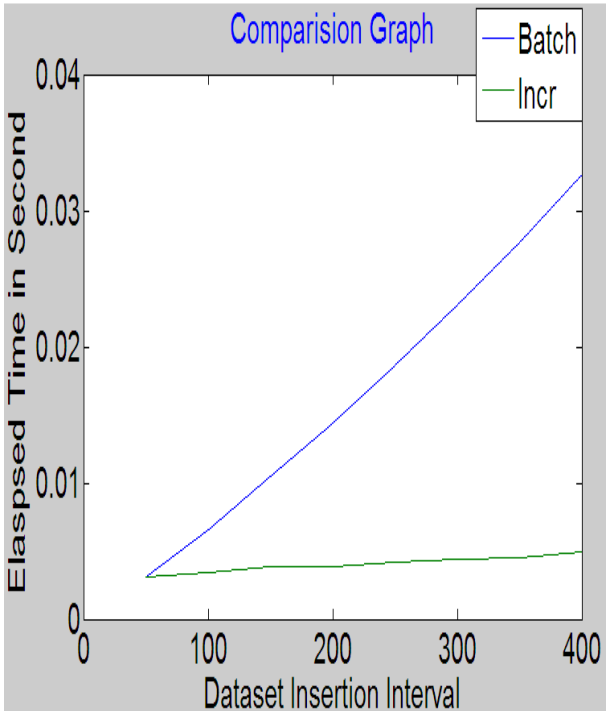

Fig. 2 Proposed work Insertion graph for Batch and Incremental Elapsed Time.

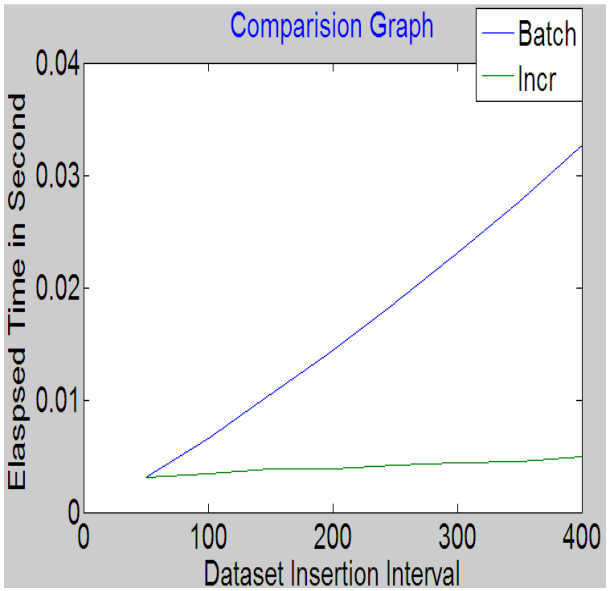

Fig. 3 Previous work Insertion graph for Batch and Incremental Elapsed

From above fig. 2 and 3 it is obtained that proposed work is better as compare to previous work in [13]. As average space required for proposed work algorithm is comparatively less. Space requirement is less because of generations of perfect patterns in proposed work. It has seen that by increase in dataset size is space cost also increases but not in same ratio as data set increases.

5. CONCLUSION

As researchers are working on different field out of which finding an effective vertical patterns is measure issue with this growing digital world. This paper has proposed an data partition algorithm for different sites. Here proper vertical patterns are generate with the help of aprior algorithm. By the use of AES encryption algorithm security of the data at server side get enhance as well. Results shows that proposed work incremental time get reduce by 9%. While batch elapsed time get reduce by 17%. By the use of automatic vertical pattern space cost is also reduce by 5%. As research is never end process so in future one can adopt other pattern generation technique for improving the server performance.

6. REFERENCES

[1] Abedjan, Z., Grütze, T., Jentzsch, A., Naumann, F.: Mining and profiling RDF data with ProLOD++. In: Proceedings of the International Conference on Data Engineering (ICDE), pp. 1198–1201(2014).

[2] Rostin, A., Albrecht, O., Bauckmann, J., Naumann, F., Leser,U.: A machine learning approach to foreign key discovery. In: Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB) (2009)

[3] Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert, Jan-Peer Rudolph, Martin Schonberg, Jakob Zwiener and Felix Naumann, "Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms", Proceedings of VLDB 2015.

[4] Huhtala, Y., Karkkainen, J., Porkka, P., and Toivonen, H., (1999), TANE: An Efficient Algorithm for discovering Functional and Approximate Dependencies, The Computer Journal, V.42, No.20, pp.100-107.

[5] Huhtala, Y., Karkkainen, J., Porkka, P., and Toivonen, Dependencies Using Partitions, IEEE ICDE 1998.

[6] Shyue-liang Wang, Jenn-Shing Tsai and Been-Chian Chien, "Mining Approximate Dependencies Using Partitions on Similarity-relation-based Fuzzy Databases", IEEE International Conference on Systems, Man and Cybernetics(SMC) 1999.

[7] Yao, H., Hamilton, H., and Butz, C., FD_Mine: Discovering Functional dependencies in a Database Using Equivalences, Canada, IEEE ICDM 2002.

[8] Wyss. C., Giannella, C., and Robertson, E. (2001), FastFDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances, Springer Berlin Heidelberg 2001.

[9] Russell, Stuart J. and Norvig, Peter. Arti cial Intelligence: A ModernApproach. Prentice Hall, 1995.

[10] Mannila, H. (2000), Theoretical Frameworks for Data Mining, ACM SIGKDD Explorations, V.1, No.2, pp.30-32.

[11] Stephane Lopes, Jean-Marc Petit, and Lotfi Lakhal, "Efficient Discovery of Functional Dependencies and Armstrong Relations", Springer 2000.

[12] Heikki Mannila and Kari-Jouko R¨aih¨a. Design by example: An application of Armstrong relations. Journal of Computer and System Sciences, 33(2):126{141, 1986.

[13] Wenfei Fan, Jianzhong Li, Nan Tang, And Wenyuan Y. "Incremental Detection Of Inconsistencies In Distributed Data". Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014 1367