International Journal of Research in Advent Technology, Vol.5, No.5, May 2017
E-ISSN: 2321-9637
Available online at www.ijrat.org

An Robust and Relative Data Distribution Approach by Utilizing Markov Model

Chitra Raghuwanshi[1], Manish Mishra[2]
Department Computer science
Email: chitra.graghuwanshi@gmail.com[1] *,Hodcs.vits@gmail.com*[2]

Abstract- The period of huge database is currently a major issue. Be that as it may, the conventional information examination will most likely be unable to deal with such extensive amounts of information. So analysts attempt to build up an elite stage to productively investigate and keep up the calculations. Here proposed work has resolve this issue of computerized information security by finding the connection between the columns of the dataset which depends on the Markov model. So columns are stored on different servers which decrease overall dataset maintained cost while security get increased. So to increase the security of information on destinations Advanced encryption calculation was utilized. Analysis is done on genuine dataset. Results demonstrates that proposed work is better as contrast with different past methodologies on the premise of assessment parameters.

Index Terms- Distributed Data, Data Mining, Encryption, Effective Pruning, Functional Dependency.

1. INTRODUCTION

Data mining methodology can help associating knowledge gaps in human understanding. Such as analysis of any student dataset gives a better student model yields better instruction, which leads to improved learning. More accurate skill diagnosis leads to better prediction of what a student knows which provides better assessment. Better assessment leads to more efficient learning overall. The main objectives of data mining in practice tend to be prediction and description [4, 5]. Predicting performance involves variables, IAT marks and assignment grades etc. in the student database to predict the unknown values. Data mining is the core process of knowledge discovery in databases. It is the process of extracting of useful patterns from the large database. In order to analyze large amount of information, the area of Knowledge Discovery in Databases (KDD) provides techniques by which the interesting patterns are extracted. Therefore, KDD utilizes methods at the cross point of machine learning, statistics and database systems.

Different approach of mining is done for different type of data such as textual, image, video, etc. Information extraction is done in digital for resolving many issues. But some time this data contain information that is not fruitful for an organization, country, raise, etc. So before extraction such kind of information is remove. By doing this privacy for such unfair information is done. This is very useful for the security of data which contain some kind of medical information about the individual, financial information of family or any class. As this make some changes on the dataset, so present information in the dataset get modify and make it general for all class or rearrange so that miner not reach to concern person.

So privacy preserving mining consist of many approaches for preserving the information at various level form the individual to the class of items [3, 4]. But vision is to find the information from the dataset by observing repeated pattern present in the fields or data which can provide information of the individual, then perturb it by different methods such as suppression, association rules, swapping, etc.

2. RELATED WORK

R.Agrawal and R.Srikant [1] utilizes ARM (Association Rule Mining) approach on large database. This paper present two algorithm based on association rule that discover relation between items. Although performance decreases with increase in database. One more point is that it does not consider item quantity information.

T.Calders and S.Verwer [2] utilizes Naive Bayes approach for classification of large database. Here author classifies dataset on the basis of frequent sensitive item sets. Here discrimination is done on the basis of gender, race, etc. which is natural class of the people. So separation done on this basis is against law, which needs to be suppressing in the dataset. Although numeric values present in the dataset remain same as previous, so it requires being perturbed as it contains many sensitive relations.

F.Kamiran and T.Calders [3] present a new approach of classification of database on the basis of non discriminating item sets. So presence of discriminating item in dataset for classification is not required. Here direct removal of sensitive information is performing. This is possible by sampling in the dataset, here

sampling make data free from discrimination. Here discriminating models are not taken for evaluation that no information is mined from operated data. But doing classification base on non discriminating items is ethical view.

In [8] multilevel privacy is provide by the author, basic concept develop in this paper is separate perturbed copy of the dataset for different user. Here user are divide into there trust level so base on the trust level dataset is perturbation percentage get increase. Here paper resolve one issue of database reconstruction by combing the different level perturbed copy then regenerate into single original database. So to overcome this problem perturbation of next level is done in perturbed copy of previous one. In this way if lower trust user get combine and try to regenerate original dataset then only one higher perturbed copy can be regenerate. The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

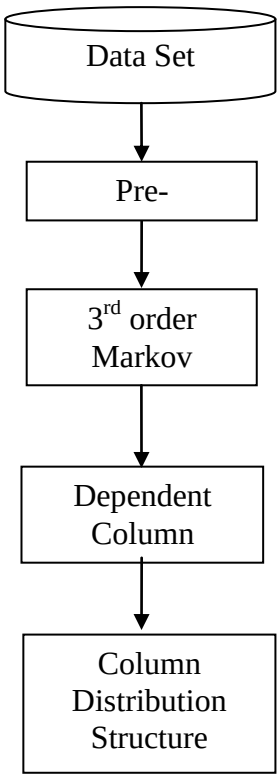
In [9, 12] paper cover a new issue for the direct indirect discrimination prevention in the dataset. Here it will collect discriminate item set which help in producing the association rule for identifying the direct or indirect rules. Then hide the rules which are above the threshold value by converting the X→Y to X→Y' where X is a set of discriminating item this tend to hide the information which will generate only those rules that not give any discriminating rule. Here Y is change to Y' means an opposite value is replace at few attributes.

3. PROPOSED WORK

Whole work is a combination of two steps where first include site creation while second include distribution of columns on various sites. While transferring whole row emcryption was performed on the them to save on the sites. Explanation of whole work is shown in fig. 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.

Kth Order Markov Modal: Let D, be a set of database transactions where each transaction T is a set of items, called Tid. Let I= {I1, I2,..., Im} be a set of items. An item set contains k items is a k item set. If a k item set satisfies minimum support (Min_sup) then it is a frequent k item set, denoted by kth markov modal. Firstly markov modal generated a set of candidates, which is candidate k-item sets,



. 1 Block diagram of proposed dependent column structure.

.

If the candidate item set satisfies minimum support then it is frequent item pattern. So base on these markov patterns columns are collect together as per large markov value.

Dependent Column
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. So if total of 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.

Column Distribution Structure
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.

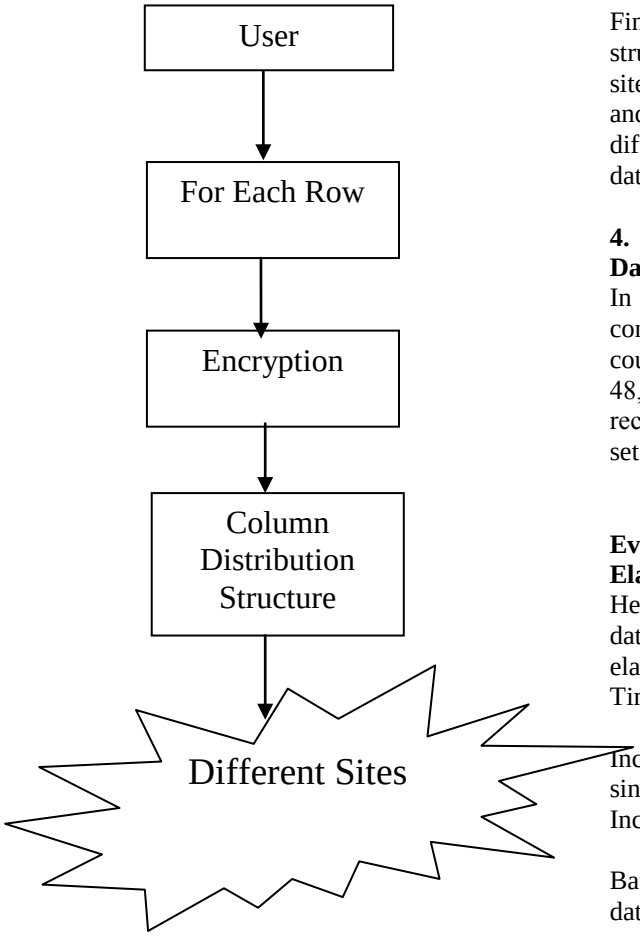


Fig. 2 User data distribution by using dependent column structure.

Insertion of Row

For each row in the data before transferring data to the selected site it need to be encrypt first. Here Advanced Encryption Algorithm is use.

AES Encryption

Now common step for all kind of data is that each data need to be convert into 16 element set of input. Here each input need to be in integer data type. In case of numeric this is ok, but in case of image gray scale will convert pixel values in integer form. While for text unique number is assign for all extracted words.

In this encryption algorithm four stages are perform in each round. While final round consist of three stages only. These steps are common in both encryption as well as decryption algorithm where decryption algorithm is inverse of the encryption one. So round consist of following four stages.

1. Substitute bytes
2. Shift rows
3. Mix Columns
4. Add Round Key

In final round simply all stages remain in same sequence except Mix Columns stage.

Finally as per the column dependencies from the structure row values are distribute between various sites. Here sites are use to maintain the pattern number and row number of the inserted row. These table at different site help in reading the required data of the dataset.

4. EXPERIMENT AND RESULT

Dataset

In [9] Sara et. al. has used 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).

Evaluation Parameters

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.

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

.

5. RESULTS

Table 3. Comparison of Average Cell Insertion time in Second.

Dataset Size	Proposed Work	Previous Work
3000	**0.00520051**	**0.0254649**
4500	**0.00321458**	**0.0184573**
6000	**0.00519351**	**0.0151151**

From above table 3 it is acquired that proposed work is better as contrast with past work in [13]. As average cell insertion time is less while executing proposed

work calculation. It has seen that by increment in dataset cell insertion also increments. As markov

model has generate patterns of columns which reduce dataset size and execution time cell insertion.

Table 4. Comparison of Average Row Insertion time in second.

Dataset Size	Proposed Work	Previous Work
3000	**0.0364035**	**0.178255**
4500	**0.0225021**	**0.129201**
6000	**0.0363546**	**0.105806**

From above table 4 it is acquired that proposed work is better as contrast with past work in [13]. As row insertion time is less while executing proposed work calculation. It has seen that by increment in dataset execution time also increments. As markov model has generate patterns of columns which reduce dataset size and execution time of row insertion is directly reduced.

Table 5 Comparison of Space Cost for data.

Dataset Size	Proposed Work	Previous Work
3000	**12745**	**37613**
4500	**17497**	**55674**
6000	**22269**	**79387**

From above table 5 it is acquired that proposed work is better as contrast with past work in [13]. As space required for dataset storage is less for proposed work calculation. It has seen that by increment in dataset space also increments. As markov model has generate patterns of columns which reduce dataset size and execution time of row insertion.

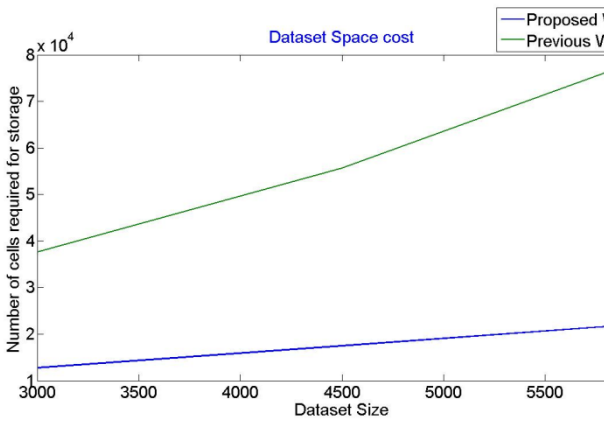


Fig. 3 Proposed work space cost at different dataset size.

From above fig. 3 it is acquired that proposed work is better as contrast with past work in [13]. As space required for dataset storage is less for proposed work calculation. It has seen that by increment in dataset space also increments. As markov model has generate patterns of columns which reduce dataset size and execution time of row insertion.

6. CONCLUSION

As scientists are chipping away at various field out of which finding a powerful vertical examples is measure issue with this becoming advanced world. This paper has proposed an information distribution algorithm for various servers. Here legitimate vertical columns are produce with the assistance of markov model. By the utilization of AES encryption calculation security of the information at server side get upgrade too. Results demonstrates that proposed work execution time get decrease. While batch passed time get decrease. By the utilization of programmed vertical example space cost is additionally diminish. As research is never end handle so in future one can embrace other example era method for enhancing the server execution.

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

[14] Thorsten Papenbrock, Felix Naumann ." A Hybrid Approach to Functional Dependency Discovery". SIGMOD'16, June 26-July 01, 2016, San Francisco, CA, USA c 2016 ACM. ISBN 978-1-4503-3531-7/16/06. .

[15] Akshay Kulkarni, Sachin Batule, Manoj Kumar Lanke, Adityakumar Gupta. "Functional Dependencies Discovery in RDBMS". International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 4, April 2016 ISSN: 2277 128X.

[16] 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