INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A ROBUST AND RELATIVE DATA DISTRIBUTION APPROACH BY SUPPRESSING ASSOCIATION RULES

[1]Nidhi Jain, [2]Prof.Angad Singh

[1]Information Technology, NRI Institute of Science and Technology, Bhopal, India
[2]Head of Department Information Technology, NRI Institute of Science and Technology, Bhopal, India
Nidhijain173@yahoo.co.in, Angada2007@gmail.com

Abstract: -The period of huge database is currently a major issue. 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 digital information security by developing the mutual understanding of different data owner for retrieving information by involving third party. Here third party generate hidden rules from the dataset in encrypted form. While each data owner suppresses association rules to a fix threshold value before sending to the server. This suppression increases the security of the information for the individual data owner. Analysis was done on genuine dataset. Results demonstrates that proposed work is better as contrast with different past methodologies on the premise of various assessment parameters.

Keywords: 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 sometime this data contains 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 providing by the author, basic concept develop in this paper is separate perturbed copy of the dataset for different user. Here user is 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 gets 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 $X \rightarrow Y$ to $X \rightarrow Y'$ 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 replacing at few attributes.

3. Proposed work

Whole work is a combination of two steps where first include site data creation (encryption, hiding association rule etc.) while second include finding patterns from the encrypted data from various data owners. Explanation of whole work is shown in fig. 1.

3.1 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 Association Rule

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. Firstly association rule generated a set of candidates, which is candidate k-item sets. If a k item set satisfies minimum support (Min_sup) then it is a frequent k item set. Firstly association rule generated a set of candidates, which is candidate k-item sets.

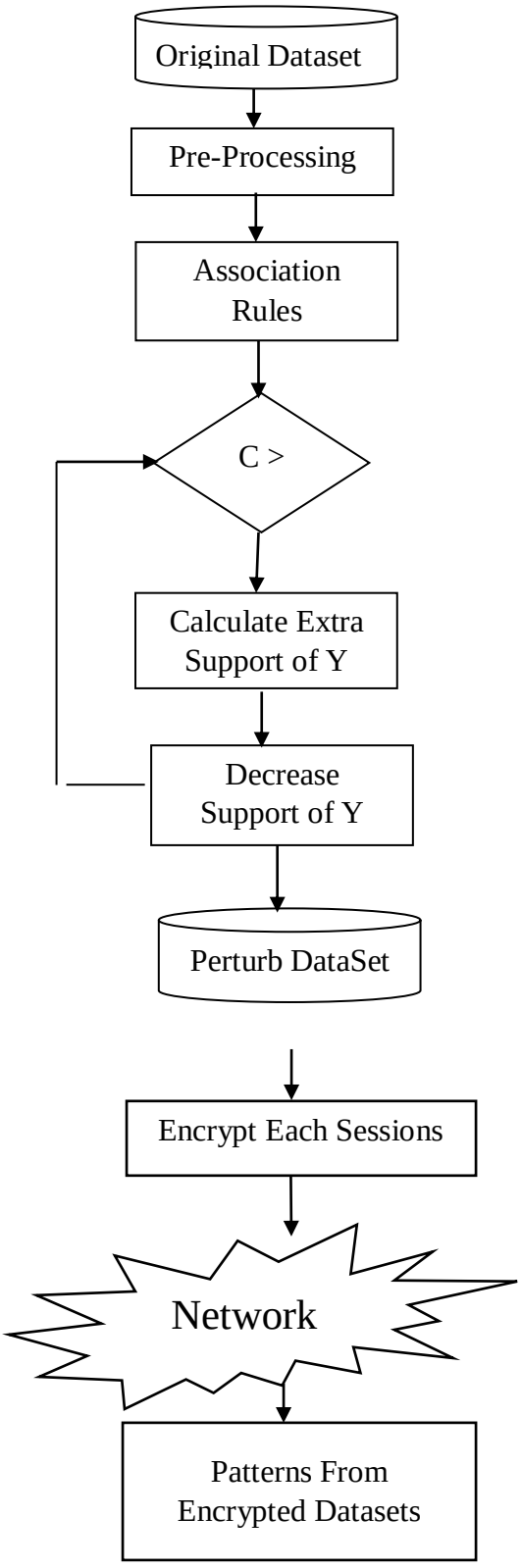

Fig. 1 Block diagram of proposed dependent column structure.

If the candidate item set satisfies minimum support, then it is frequent item pattern. So base on these association rules sensitive rules are identified.

3.3 Hide Sensitive Pattern

So in order to hide an pattern, {X, Y), it can decrease its support to be smaller than user-specified minimum support transaction (MST). To decrease the support of a rule, there is a approach: Decrease the support of the item set {X, Y}. For this case, by only decrease the support of Y, the right hand side of the rule, it would reduce the support faster than simply reducing the support of {X, Y}.

P= ((Rule support – Minimum_ support) * Total transaction)

Above formula specify the number of transaction where one can modify and overall support of that hiding pattern is lower than the minimum support.

3.4 Decrease Rule Support

In this work once number of transactions to be perturb is calculate than Y elements of the rule is replaced by F element. Here F is an artificial object introduce to reduce the support value of the rule. Although replacement of element Y1 by F1 or Y2 by F2 is fix. Therefore, if all P transaction which contain both X, Y element are replaced by F than support value of the rule is no greater than minimum support value**.**

3.5 ***Pailler cryptosystem****:*

This cryptosystem is based on the public and private key concept. Here input vector D[n], will be encrypt by this algorithm.

1. Choose two large prime numbers p and q randomly and independently of each other such that gcd(pq,(p−1)(q−1))=1.
2. Compute RSA modulus $n = pq$ and
   Carmichael's function $\lambda\lambda=llllll(pp-1,qq-1)$
3. Select generator g, Select α and β randomly from a set $\mathbb{Z}n*n$ then calculate $g=(\alpha n+1)\beta*\beta \bmod(n*n)$
4. Calculate the following modular multiplicative inverse
   $\mu=\bmod(n) / (L(g\lambda \bmod(n*n)-1$

   Where the function L is defined as $(u)=(u-1)/n$.

So The public key is (n, g), private key is (λ, μ).

4. Patterns from Encrypted Datasets

In order to generate patterns from the different encrypted datasets of the various users each column from the datasets are combine into single one for developing a single table. Here based on the different numeric value of the column patterns are generated where each pattern are count in whole dataset. Here patterns are generating from column data obtaining from different data owner. It means same data owner column are not consider for finding the rules as it is assumed that data can himself find that pattern. This can be understood by below example.

Table 1. Dataset obtained from different data owners.

Data Owner 1			**Data Owner 2**	
Column1	**Column2**		**Column1**	**Column2**
A1	B1		X1	Y1
A2	B2		X2	Y2

Table 2. Merge dataset obtained from different data owners.

Column1	**Column2**	**Column3**	**Column4**
A1	B1	X1	Y1
A2	B2	X2	Y2

Now rules generate from the above dataset are A1→X1, A1→X2, A1→Y1, A1→Y2, B1→X1, B1→X2, B1→Y1, B1→Y2, etc. So support of each rule is calculate and send to each data owner how are participating in this pool. Finally, each data owner decrypt the values for getting exact name of the object.

5. Experiment

5.1. Dataset

In [9] Sara et. al. has used Adult dataset where it contains 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).

5.2. Evaluation Parameters

5.2.1 Elapsed Time

Here total execution time (second) is calculate for the dataset which was required to find the rules from different data owners.

5.2.2. Rule Count

Here number of association rules are count which was generate by the server from the different data owner's datasets. If large number of rules are generate due to fake data than confusion for the correct information is more.

5.2.3. 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.

6. Results

Table 3. Comparison of elapsed time in Seconds.

Dataset Size	Proposed Work	Previous Work
500	0.0665	0.1707
1000	0.1134	0.2089
2000	0.1984	0.4391
4000	0.3385	0.6325

From above table 3 it is acquired that proposed work is better as contrast with past work in [13]. As elapsed time is less while executing proposed work calculation. It has seen that by increment in dataset cell elapsed time also increments. As fake transactions increase the dataset size so execution time of previous method was high.

Table 4. Comparison of rule count.

Dataset Size	Proposed Work	Previous Work
500	56	94
1000	64	106
2000	84	124
4000	94	126

From above table 4 it is acquired that proposed work is better as contrast with past work in [13]. As rule count at server is less by the proposed work calculation, this shows that number of false rules are less as compare to previous approach. It has seen that by replacing the fake elements in the dataset fake rules are obtained at the server side which reduce the true rule count of the proposed work.

Table 5 Comparison of Space Cost for data.

Dataset Size	Proposed Work	Previous Work
500	500	565
1000	1000	1130
2000	2000	2260
4000	4000	4520

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.

7. 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 secured information distribution algorithm. By the utilization of Paillers encryption calculation security of the information at server side get

increases. Results demonstrates that proposed work execution time get decrease. 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 (Web DB) (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), Fast FDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances, Springer Berlin Heidelberg 2001.

[9] Russell, Stuart J. and Norvig, Peter. Artificial Intelligence: A Modern Approach. 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, Aditya Kumar 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