Volume 2, Issue 5, May 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Approaches of Privacy Preservation in Data Mining

Ashwini Niranjan Ingle[1]
M.E., Dept. of CSE
Prof. Ram Meghe Institute of Technology & Research
Badnera, India

Prof. S. V. Pattalwar[2]
Associate Professor, Dept. of CSE
Prof Ram Meghe Institute of Technology & Research
Badnera-Amravati, India

Prof. P. K. Agrawal[3]
Assistant Professor, Dept. of CSE
Prof Ram Meghe Institute of Technology & Research
Badnera-Amravati, India

Abstract: ***In the decade data mining has lured attention in information industry. Data mining deals with extraction of useful and important data from the collection of huge data. We all have information on our finger tips but data mining helps in retrying the useful information. Information is the important and crucial data which is to be kept secured. This paper has dealt with various security challenges, issues and the approaches for preserving privacy in data mining. Preserving privacy and providing security are both different. When two organization or people deals to share their data.***

Keywords: ***Security Issues, Cloud Security, Cloud Architecture, Cloud Platform.***

I. INTRODUCTION

Data mining deals with the information which is being extracted from the data. When information is said as data it possess some substantial value. Data mining have different applications in various domains such as business, medical, gaming application, web services, counter terrarium application etc . Due to its usefulness and popularity it has been used widely. But as it is widely used privacy and security has also became a matter of concern . privacy and security are both different aspects.

Now a days, internet has become a source for sharing information and also a resource of data storage and retrieval which has risen the need of privacy of data. There are various approaches of privacy preservation in data mining. There are various places where the information is to kept confidential such as medical history, contract numbers, banking details, income etc. Not always the shared information may incur loss to the individual or to the organization. But if the information, when hared between two organization viewed by the third party it may be misused or modified. Once the information is revealed it is hard to retrieve the information or the loss caused due to it. Privacy is not violated till one feels his personal information is being used negatively[4]. The goal of PPDM is to develop algorithms and techniques by which original data is modified in some way that mining process cannot extract original confidential data. Privacy preserving data mining has becoming an increasingly emerging topic of research [4].

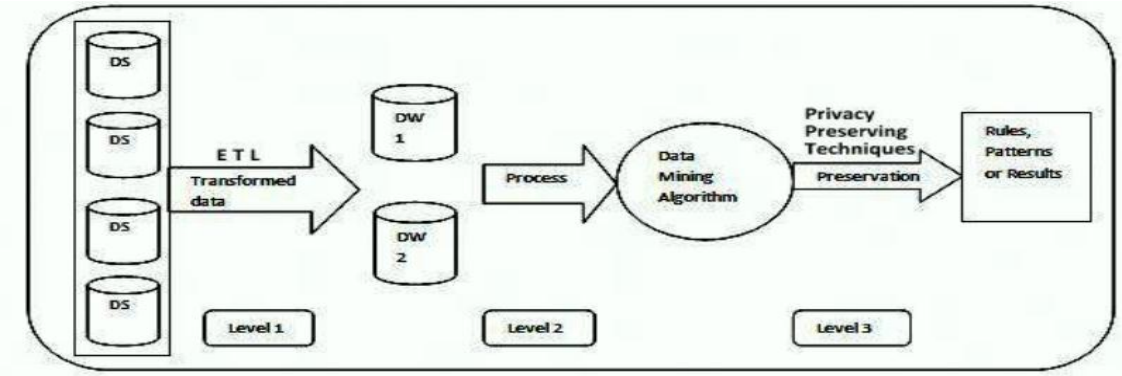

Fig. 1 Framework of privacy preserving data mining[4]

In this paper, the section II gives Literature Survey, section III describes privacy preserving methods, section IV gives advantages, V the proposed work for system. Followed by future scope and conclusions summarized.

II. CLOUD DEPLOYMENT MODEL

- Secure two party computation was first investigated by Yao [10],and was later generalized to multi-party computation in [11,12,13].
- Privacy preserving[6] data mining (PPDM) protects privacy of the data of person without letting go the the usefulness of the data. Many method are proposed for the same.
- T- Closeness model [7] uses the k-anonymity and 1 diversity approach but in addition ensures that that the distances between the distributions of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. Compared to the previous methods. This model provides better privacy but there is information gain for the attacker and data characteristics are also lost.
- In [8] R. Agrawal and R.Srikant started the work towards PPDM. They categorized the methods as perturbation and secure multi-party computation. Many variation of algorithm has been suggested like database extension geometric data. X. Xiao, Tao, and M. Chen introduces the multiple version of the dataset and anonymized data set at different privacy levels.
- G. Wang. z.zhu, W. DU, and Z.Teng proposed maximum Entropy Principle to perform interface analtsis for disguised dataset[9].
- In [17]Author has proposed the privacy preserving data mining technique in Hadoop that. solve privacy violation without utility degradation but its execution time affected by noise size.
- In [18] a methodology has been proposed which provides data confidentiality, secure data sharing without Re-encryption, access control for malicious insiders, and forward and backward access control.

III. PRIVACY PRESERVING TECHNIQUES

There are many methodologies which have been accepted for privacy preserving data mining. We can categorize them based on the following measurements:[14]

•Data Distribution

•Data Modification

•Data Mining Algorithm

•Data or Rule hiding

•Privacy Preservation

Data distribution deals with distribution of data. There are centralized data and distributed data, approaches are developed for both centralized as well as distributed data. Distributed data are further categorised into vertical data distribution and horizontal data distribution.

Data values in the database is being modified for ensuring high privacy protection[15][16].data modification technique should be in concert with the privacy policy. There are several data modificationn technique[14].

a. Perturbation: which is able to replacing attribute value by a new value (changing a 1-value to a 0- value, or adding noise)

b. Blocking: which is the replacement of an existing attribute value with a "?"

c. Swapping: This refers to interchanging values of individual record.

d. Sampling: This refers to losing data for only sample of a population.

e. Encryption: many Cryptographic techniques are used for encryption.

The third and important point discusses the data mining algorithm, all these distribution and modification is done for mining of data only. There are various data mining algorithms. Which helps to analayze and retrieve the information from data. It also helps in data hiding.

Widely used algorithm of mining for the purpose classification Bayesian networks, association rule mining algorithms, decision tree,etc.. data mining algorithm are[14]:

1. Classification data mining algorithm
2. Association Rule mining algorithms
3. Clustering algorithm

Fourth category deals with data hiding. In case of data sharing when large amount of information is shared there is risk and less information is shared it affects the data miner. The complexity at this phase is high.

The last is privacy preservation which is used for selective modification of the data. For higher utility of the modified data and that the policy is not jeopardized[14]. Privacy Preserving Techniques:[19]

1. Heuristic-based techniques: It is an adaptive modification that modifies only selected values that minimize the effectiveness loss rather than all available values.

2. Cryptography-based techniques: This technique includes secure multiparty computation where a computation is secure if at the completion of the computation, no one can know anything except its own input and the results. Cryptography-based algorithms are considered for protective privacy in a distributed situation by using encryption techniques.

3. Reconstruction-based techniques: where the original distribution of the data is reassembled from the randomized data.

Based on these dimensions, different PPDM techniques may be classified into following five categories [20-22, 23, 24].

a. Anonymization based PPDM
b. Perturbation based PPDM
c. Randomized Response based PPDM
4. Condensation approach based PPD

IV. ADVANTAGE AND LIMITATIONS

Table 1. Advantages and Limitations of PPDM Techniques[19]

Technique	Advantages	Limitations
Anonymization based PPDM	Identity or sensitive data about record owners are to be hidden.	Linking attack. Heavy loss of information.
Perturbation based PPDM	In this technique different attributes are preserved independently.	Original data values cannot be regenerated. Loss of information.
Randomized Response based PPDM	It is relatively simple useful for hiding information about individuals. Better efficiency compare to cryptography based PPDM technique [20].	Loss of individual's information. This method is not for multiple attribute databases.
Condensation Approach based PPDM	Use pseudo data rather than altered data. This method is very real in case of stream data.	Huge amount of information lost. It contain same format as the original data.
Cryptography based PPDM	Transformed data are exact and protected. Better privacy compare to randomized approach [20].	This approach is especially difficult to scale multiple parties are involved.

V. PROPOSED METHOD

In data mining many techniques are there for providing privacy to the data. Using hybrid can give better result. In[24] author has used genetic algorithm and PSO. Genetic algorithm when implemented with association rules[1] could provide better results. Using hybrid approach It makes difficult for the attacker to identify background and homogeneity attack.Result will show fairly good level of privacy.

Genetic Algorithm (GA)

In Genetic Algorithm (GA), a group of individuals called chromosomes forms the population that represents a complete solution to a defined problem [27, 28].

1. Crossover: The offspring is generated from two chosen individuals in the population by swapping some attributes among the two individuals. The offspring inherits some characteristics from both of the two individuals (parents).
2. Mutation: Mutation: One or several attributes of an offspring may be randomly changed. The offspring may possess different characteristics from their parents. Mutation increases the possibility of achieving global optimization.
3. Selection: Excellent offspring are chosen for survival according to predefined rules. This operation keeps the population size within a fixed amount, and the good offspring have a higher possibility of getting into the next generation.

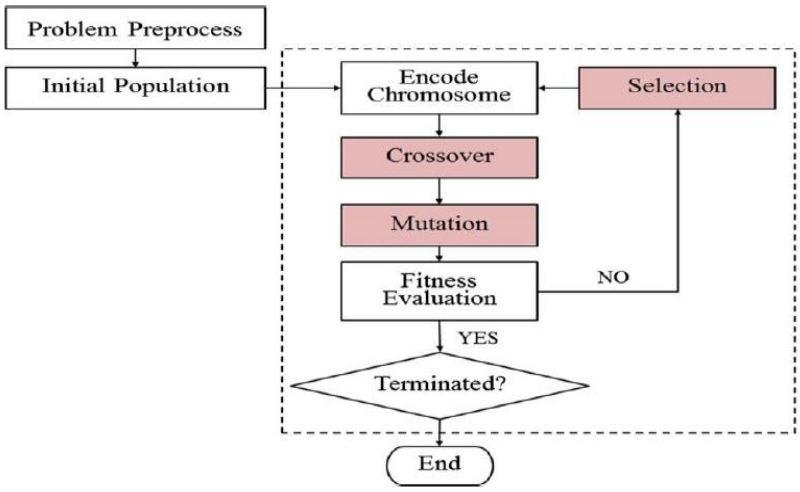

Fig. 2 Flowchart of GA[4].

CONCLUSION

Using hybrid approach It makes difficult for the attacker to identify background and homogeneity attack Implementation of the proposed can be done as the future work by taking various evaluation parameters such as performance, data utility, uncertainity level, resistance etc. Apart from that it protects private data with better accuracy and gives no loss of information which increases data utility. Result will show fairly good level of privacy.

REFERENCE

1. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.
2. L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining," in Proc. Austral. Inst. Comput. Ethics Conf., 1999,
3. pp. 89_99.
4. R. Agrawal and R. Srikant, ``Privacy-preserving data mining," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439_450, 2000.
5. Deepa Tiwari, Raj Gaurang Tiwari "A Survey on Privacy Preserving Data Mining Techniques" IOSR Journal of Computer Engineering (IOSR-JCE)
6. e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 5, Ver. III (Sep. – Oct. 2015), PP 60-64
7. Halak P. PatelWarish D. Patel "A Hybrid Approach for Privacy Preserving using Randomization for Data mining" Vol-2 Issue-3 2016 IJARIIE-ISSN(O)-2395-4396
8. M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.
9. "An Efficient Approach for Privacy Preserving in Data" Mining"Manish Shannal Atul Chaudhar/ Manish Mathuria3 Shalini Chaudhar/ Santosh Kumar5, IEEE , 2014
10. "Data Anonymization using Augmented Rotation of sub-cluster for Privacy preservation in data mining"V.Rajalakshmi, G.S.Anandha Mala,IEEE,2013
11. "A Combine Random noise perturbation approach for multi-level privacy preservation in data mining" Mr.S.Chidambaram, Research Scholar/NEC,Dr.k.g Srinivasagam/CSE/NEC,IEEE,2014
12. A.C. Yao, How to generate and exchange secrets, Proc. of the 27th IEEE Symp. on Foundations of Computer Science, 1986, pp. 162–167.
13. M. Bellare and S. Micali, *Non-interactive oblivious transfer and applications*, Advances in Cryptology - Crypto '89, pp. 547-557, 1990.
14. M. Ben-Or, S. Goldwasser and A. Wigderson, *Completeness theorems for non cryptographic fault tolerant distributed computation*, 20th STOC, (1988), 1-9.
15. D. Chaum, C. Crepeau and I. Damgard, Multiparty unconditionally secure protocols, 20th Proc. ACM Symp. on Theory of Computing, (1988), 11-19.

16. Md. Riyazuddin, Dr.V.V.S.S.S.Balaram , Md.Afroze, Md.JaffarSadiq, M.D.Zuber ,"An Empirical Study on Privacy Preserving Data Mining", International Journal of Engineering Trends and Technology- Volume3Issue6- 2012
17. Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, Disclosure Limitation of Sensitive Rules, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999),45–52.
18. Chris Clifton and Donald Marks, *Security and privacy implications of data mining*, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.
19. "Hiding a needle in a Haystack: privacy preserving Apriori algorithm in map reduce framework" ACM Nov 7, 2014
20. Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.View ArticleGoogle Scholar
21. Hina Vaghashia, Amit Ganatra," A Survey: Privacy Preservation Techniques in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015.
22. S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004, pp: 50-57.
23. Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.
24. Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.
25. Dharmendra Thakur and Prof. Hitesh Gupta," An Exemplary Study of Privacy Preserving Association Rule Mining Techniques", P.C.S.T., BHOPAL C.S Dept, P.C.S.T., BHOPAL India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 11,2013.
26. C.V.Nithya and A.Jeyasree,"Privacy Preserving Using Direct and Indirect Discrimination Rule Method", Vivekanandha College of Technology for WomenNamakkal India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 12,2013.
27. Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi," A Hybrid Algorithm for Privacy Preserving in Data Mining", I.J. Intelligent Systems and Applications, 2013, 08, 47-53
28. Narges Jamshidian Ghalehsefidi1, Mohammad Naderi Dehkordi," A Hybrid Approach to Privacy Preserving in Association Rules Mining", ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 6, No.12 , November 2014 ISSN : 2322-5157
29. L. David, Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold. 1991.
30. D.E. Goldberg, Genetic Algorithms: in Search, Optimization, and Machine Learning. New York: Addison-Wesley Publishing Co. Inc. 1989.