Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN : 2348-4470
p-ISSN : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3,Issue 5,May -2016

Feature Selection in Privacy Preserving in Data Mining-Review

Kavisha Patel[1], Amit Rathod[2]

[1]Department of Computer Engineering, PIET, Vadodara,
[2]Department of Information Technology, PIET, Vadodara,

Abstract- A key problem that arises in any mass collection of data is that of confidentiality of the data. Privacy-preserving data mining (PPDM) is the area of mining that seeks to safeguard sensitive information from uninvited or unofficial speech act of individual information records. There are several basic techniques like anonymization, cryptography and randomization. The attributes are segregated supported their sensitivity for privacy preservation purposes. There are many basic privacy preservation techniques like anonymization, cryptography and randomization. The attributes are segregated based on their sensitivity for privacy preservation purposes. Automating this attribute segregation becomes complicated in high dimensional datasets and data streams. Information or correlation of the attribute on the target class attribute is measured using Information Gain [IG], Gain Ratio [GR] and Pearson Correlation [PC] ranker based feature selection methods with decision tree and this values are used to segregate them as Sensitive Attributes [SA], Quasi Identifiers [QI] and Non-Sensitive. Segregated attributes are subjected to various levels of privacy preservation using both the Double layer Perturbation [DLP] and Single Layer Perturbation [SLP] algorithms to form the level-1 perturbed datasets. Since the attribute selection uses tree structure, the work proposes a linked array instead of tree to reduce the number of iterations and increase the efficiency.

Keywords- privacy preserving, feature selection, perturbation

I. Introduction

Protecting user personal knowledge is a crucial concern for society. The daily use of the word privacy concerning secure data sharing and analysis is commonly imprecise and should be dishonest. Privacy protective started within the Nineteenth Sixties once variety of researchers recognized the problem of privacy violations by massive collections of non- public data in laptop systems[1].
To protect privacy of individual, many strategies will be applied on knowledge before or on the method of mining. The branch of study that embody these privacy considerations are referred as Privacy Preserving Data Mining(PPDM). Privacy will be achieved through any one of the strategies like knowledge hiding/masking, suppression, generalization, anonymization, perturbation, encryption, organization, condensation, fuzzification, secure multi-party computation, etc.
Over the years variety of definitions in privacy protective has emerged. one in all them defines "privacy protective because the every individual's ability to manage the circulation of data concerning him/her". it's additionally outlined as "the claim of single person, groups, or establishments to work out for themselves once, however & what extent the data concerning them is communication with others". Privacy is additionally concerned confidentiality and security of the info and over two[1].

II. Literature Survey

1. Privacy Preserving Data Mining Techniques-Survey, Ms. Dhanalakshmi M and Mrs. Siva Sankari E, IEEE, 2014[1]

In this paper, introduction to privacy preserving is provided. It provides classification of privacy preserving technique based on different dimensions- Data distribution, Data modification, Data mining algorithm, Data or rule hiding.
There is a detailed survey done on the preservation techniques like Randomization, Anonymization and Encryption methods.

- Randomization Method

The randomization method provides an efficient however easy approach of preventing the user privacy from learning sensitive data, which might be simply enforced at information assortment section for privacy preserving Data mining. During this methodology the extra information else to a given record is freelance of the behavior of different original information records. Once the randomization method is administered, the info assortment method consists of 2 steps. Representative randomization methods embrace random-noise-based perturbation and randomised response theme. The randomization method may be a easy technique which might be simply enforced at information assortment time. It's been shown to be a helpful technique for concealing individual information in privacy preserving data mining. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining. The randomization method is additional economical .

- Anonymization Method

Anonymization method aims at creating the individual record be indistinguishable among a bunch record by victimisation techniques of generalization and suppression.

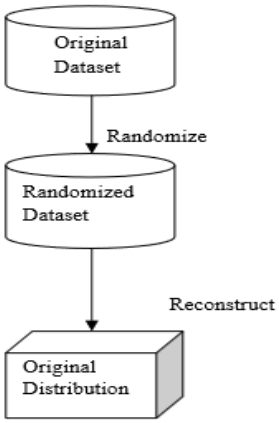

Fig. 1 Model of Randomization Method[1]

Many advanced methods have been projected, like p-sensitive k- anonymity, M-invariance, Personalized anonymity, (a, k) -anonymity, l-diversity, t-closeness and so on. The anonymization method make sure that the reworked information is correct, however it conjointly leads to data loss in some extent.

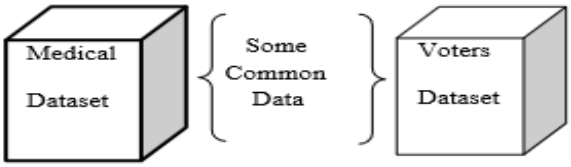

Fig. 2 Example of Anonymization Method[1]

- Encryption Method

Encryption method chiefly resolves the issues that individuals put together conduct mining tasks supported the non-public inputs they supply. These privacy mining tasks may occur between mutual un-trusted parties, or maybe between competitors. Therefore, to guard the privacy becomes a crucial concern in distributed data processing setting. All the strategies are nearly supported on the special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC technology used in distributed privacy preserving data mining areas mainly consists of a set of secure sub-protocols, such as, secure sum, secure intersection, secure set union, secure comparison, dot product protocol and so on.

2. A Survey on Anonymity-based Privacy Preserving, Jian Wang, Yongcheng Luo, Shuo Jiang, Jiajin Le, IEEE, 2009[2]

In this paper, the technology of has been planned to safeguard sensitive attributes from the corresponding identifiers. This paper intends to tell many anonymity- based mostly privacy protective technologies clearly. They need first shown that a k-anonymity dataset permits sturdy attacks as a result of lack of diversity within the sensitive attributes. Then they need shown l-diversity, a framework that offers stronger privacy guarantees. K- anonymity and l-diversity are studied wide as mechanisms for preventing re-identification attacks in micro-data unleash. Then they need shown a straight forward and effective privacy preservation technology referred to as Anatomy. ultimately they need shown a brand new privacy live t-closeness. Besides, they additionally analyze the deserves and shortcomings of those technologies.

3. Gain Ratio based Feature selection method for privacy preservation r. Praveena Priyadarsini, M.L. Valarmathi and S. Sivakumari, ICTACT Journal on soft computing, 2011[3]

In this paper, k-anonymity privacy protection technique is applied to high dimensional datasets like adult and census. Since, each the information sets square measure high dimensional, feature subset selection method like Gain Ratio is applied and also the attributes of the information sets square measure hierarchical and low ranking attributes square measure filtered to create new reduced data subsets [7]. K-anonymization privacy preservation technique is then applied on reduced datasets. The k- anonymized original and reduced square measure compared for accuracy on each data processing task classification and clustering. The obtained results that showed the accuracy level remained constant for

k-anonymized original datasets and reduced datasets for each data mining functionalities . This shows that the utility of each the datasets aren't stricken by each spatial property reduction and privacy preservation K- anonymization technique.

4. Attribute Segregation based on Feature Ranking Framework for Privacy Preserving Data Mining, R. Praveena Priyadarsini1, M. L. Valarmathi and S. Sivakumari, Indian Journal of Science and Technology, 2015[4]

In this paper, information or correlation of the attribute on the target class attribute is measured using Information Gain [IG], Gain Ratio [GR] and Pearson Correlation [PC] ranker based feature selection methods and this values are used to segregate them as Sensitive Attributes [SA], Quasi Identifiers [QI] and Non-Sensitive [NS] Attributes. Segregated attributes are subjected to various levels of privacy preservation using both the proposed Double layer Perturbation [DLP] and Single Layer Perturbation [SLP] algorithms to form the level-1 perturbed datasets[8]. The level-1 perturbed dataset is further perturbed by applying SLP algorithm to form level-2 and level-3 privacy preserved datasets. Thus, the multiple versions of Adult information set created square measure distributed to data seekers supported their trust levels in Multi Trust Level [MTL] environment[8]. The privacy preserved dataset versions created victimization the planned algorithms square measure evaluated supported their utility, distortion and purity metrics. The results show that the ranker methods are able to identify attributes which had sensitive content as either SA or QI automatically. Also, once perturbated versions of datasets square measure evaluated supported distortion metrics, all the perturbated versions have sensible distortion values so preventing diversity attacks. in comparison for its utility with the first dataset and L-Diversified Adult dataset there's a awfully tiny variation in accuracy altogether the planned perturbated datasets. Thus, the experiments show that perturbation of high graded attributes doesn't have abundant impact on the utility of the datasets.

5. Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining, Praveena Priyadarsini, M.L. Valarmathi, S.Sivakumari, International Journal of Computer Applications, 2012[5]

In this paper privacy protection is applied to high dimensional datasets like Adult and Census. For ranking the attributes, info gain feature set choice technique is employed. The high ranking attributes with sensitive info square measure set as similar identifiers of the datasets. A hybrid perturbation technique is employed to perturb categorical and numeric attributes of each the datasets and also the utility of the datasets is measured exploitation accuracy on data processing functionalities. The info distortion is measured exploitation using maintenance of Rank of Features (CK) between the original and perturb datasets. the initial and perturb datasets. The results of the projected perturbation techniques compared with alternative privacy preservation techniques like k- obscurity and l-diversity. The results showed that the extent of accuracy and thus the utility remained a similar for the each original and privacy preserved datasets. The prognosticative accuracy of the the projected flustered dataset is comparable alternative techniques. Once the upkeep of Rank of Features (CK) live of similar identifiers in each the datasets is compared, Census dataset has higher CK price than Adult dataset since the attribute rank price of sophistication of employee attribute in Census dataset hyperbolic when perturbation.

III. CONCLUSION

Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. The survey of various privacy preserving techniques is done with each having advantages and disadvantages over different parameters. Privacy techniques should be mainly provided to the sensitive attributes. This attribute selection is done by using different attribute selection measures like Information Gain, Gain Ratio, etc. which uses tree structure and provide rank value which classifies sensitive data and other non-sensitive data.
In Future, the use of linked array can be done instead of tree structure that can reduces the iteration that occurs in tree structure and comparison is done between both the system.

REFERENCES

[1] Dhanalakshmi, M., and E. Siva Sankari. "Privacy preserving data mining techniques-survey." Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014.
[2] Wang, Jian, et al. "A survey on anonymity-based privacy preserving." E-Business and Information System Security, 2009. EBISS'09. International Conference on. IEEE, 2009.
[3] Priyadarsini, R. Praveena, M. L. Valarmathi, and S. Sivakumari. "Gain Ratio Based Feature Selection Method For Privacy Preservation.", ICTACT Journal On Soft Computing, April 2011
[4] Priyadarsini, R. Praveena, M. L. Valarmathi, and S. Sivakumari. "Attribute Segregation based on Feature Ranking Framework for Privacy Preserving Data Mining." Indian Journal of Science and Technology 8.17 (2015).
[5] Priyadarsini, Praveena, M. L. Valarmathi, and S. Sivakumari. "Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining." International Journal of Computer Applications 58.2 (2012): 34-41.
[6] Han J, Micheline K, Pei J. Data mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann; 2011.

[7] Lin, Pengpeng, et al. "Feature Selection: A Preprocess for Data Perturbation." IAENG International Journal of Computer Science 38.2 (2011): 168-175.
[8] Lin P. A comparative study on data perturbation with feature selection. Proceedings of International Multi Conference of Engineers and Computer Scientist (IMECS'2011); 2011 Mar. p. 1–168.
[9] Frank A, Asuncion A. UCI machine learning repository. Irvine, CA: School of Information and Computer Science, University of California; 2010. Available from: http://archive.ics.uci.edu/ml
[10] Mozafari B, Zaniolo C. Publishing naive Bayesian Classifiers: Privacy without accuracy loss. 35th International Conference on Very Large Data Bases (VLDB); 2009.
[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter, Vol.11, No.1, pp. 10-18, 2009.

