A Survey Paper on an Integrated Approach for Privacy Preserving in High Dimensional Data using Randomized and SVD Algorithm

Tripti Singh Thakur[1] Dr. Abha Choubey[2]
[1]P.G. Scholar [2]Associate Professor
[1,2]Department of Computer Science & Engineering
[1,2]SSGI, SSTC, Bhilai(C.G.)

Abstract— Data mining is a technique which is used for extraction of knowledge and information from large amount of data collected by hospitals, government and individuals. The term data mining is also referred as knowledge mining from databases. The major challenge in data mining is ensuring security and privacy of data in databases, because data sharing is common at organizational level. The data in databases comes from a number of sources like – medical, financial, library, marketing, shopping record etc so it is foremost task for anyone to keep secure that data. The objective is to achieve fully privacy preserved data without affecting the data utility in databases. i.e. how data is used or transferred between organizations so that data integrity remains in database but sensitive and confidential data is preserved. This paper presents a brief study about different PPDM techniques like- Randomization, perturbation, Slicing, summarization etc. by use of which the data privacy can be preserved. The technique for which the best computational and theoretical outcome is achieved is chosen for privacy preserving in high dimensional data.

Key words: Randomization; slicing; perturbation; summarization; data integrity

I. Introduction

Data mining deals with extraction of hidden confidential information from large data bases. For the process of picking out relevant information some techniques like-clustering, classification, associations are used. The concept of privacy preserving data mining means preserving personal information from data mining algorithms. For privacy preservation of data sometimes data is altered before delivery and sometimes it is altered after delivery (before showing it to the third party) . Now days we have different data mining applications for which security is a must for example- stock market, finance etc. In data mining system there are so many privacy preserving methods. The objective of privacy preserving data mining is to search some technique in which the original information is transformed in some way so that the private data and private knowledge remain confidential after the mining process.

Data cleansing only takes effect on certain type of errors and cannot result in perfect data, eliminating noisy data may lead to information loss. Noise corrupted data can be modified by the use of noise [1] knowledge. Data privacy can also be preserved by random non-linear data transformation [4].SVM classifier is also applicable for privacy preservation problems [5].WCNN algo is improved SVD method which perform data perturbation, which is more efficient in balancing data privacy and utility [6]. Enabling multilevel trust poses new challenges for perturbation based PPDM. The key challenge lies in preventing the data miners from combining the copies at different trust levels to jointly reconstruct the original data more accurate [8]. Privacy can also be preserved by decision tree learning on unrealized data sets [9]. Sample selection maintain ration of data between mixed data and original data. The singular value decomposition prevents the loss of mixed data and extracted data in decomposition of matrix [10]. Because of the increasing capability to trace and gather large amount of sensitive information privacy preserving in data mining applications has become an important concern[15].If data transformation and encryption techniques are applied in combination then the data privacy is preserved strongly [17].In Vector Quantization after encoding one can not reveal the original data hence privacy is preserved. The vector quantization methods are efficiently applied in the development of speech recognition system [20].

II. Problem definition

Due to the increase in sensitive information in databases privacy preservation is an important concern for each and every data miner. The need for privacy is sometimes due to law(for medical databases) or can be influenced by medical interest. For scientific, economic and market oriented databases confidentiality is an important issue. So it is important to develop such a technique for privacy preservation that the data utility and integrity remains constant without affecting the data confidentiality.

III. Methodology

Main goal of privacy preserving data mining is to detect such solution which will result data security with data integrity with low computational complexity. We have the following different techniques for privacy preservation data mining which have been proposed after deep study of data mining community-

A. Randomization Technique

In randomization technique the original data is hidden by randomly modifying the data values. It provides some deeper statistical approach to security and privacy. On dataset randomization is applied either by adding or by multiplying random values to original records. Randomization technique is Easy to implement and it has a very High search accuracy. It is computationally efficient and Suitable for different user requirements. But it results High information loss.

B. Perturbation Technique

This technique is also known as noise addition technique. Here noise is introduced either to the data or to the result of the queries. Perturbation is done by use of Gaussian and uniform perturbing functions. Here each and every attribute is treated independently but the Data Confidentiality has been compromised sometimes.

C. Cryptographic Technique

In cryptographic technique data value is altered by some encryption technique like secure sum, secure set union, and secure size set intersection etc. on all data or only on confidential data values. The output of a computation is not protected by cryptographic techniques instead it prevents privacy leaks during computation. It provide High search Accuracy but with high Computational complexity.

D. Slicing

In this technique the data set is partitioned both vertically and horizontally. The basic idea of slicing is to preserve the association within each column and break association between uncorrelated attributes, which are infrequent. Here reduced data dimensionality is achieved with preserved data utility. It preserves better utility in micro data but It is Not used for high dimensional data.

E. Summarization

In this technique the data is released in the form of a “summery” which only having some aggregate queries not the confidential records. Summarization is applied either by sampling or by tabular representation of data values. It is mostly used for tabular data but privacy preservation is not guaranteed.

F. Suppression

In this technique suppression is applied on sensitive data before any computation. Use of suppression increases the data utility and reduces the amount of sensitive data on dataset. Here protection from discovery of certain statistical characteristics, such as sensitive association rules is done.

G. SVD Technique

It is a matrix factorization technique which is most widely used in PPDM. It perturbs every sample of data to the same degree and Provide less information loss but for different user requirements it is not suitable.

IV. Result

After studying different privacy preserving data mining techniques the result we conclude is that If randomization is used with perturbation technique and SVD technique Better privacy may be achieved with increased data utility because In the combination of these three techniques the cons of each techniques are recovered by the pros of other two techniques. The graphical representation of expected outcome is shown below

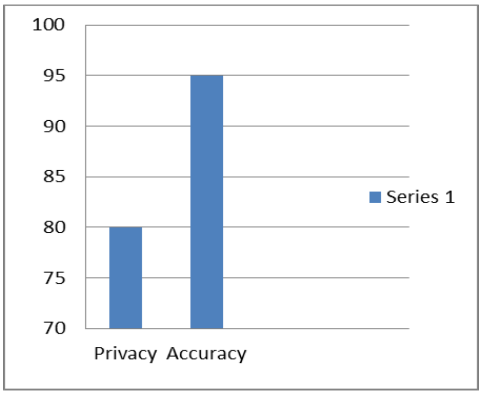

Fig. 1: Graph .1: Graphical representation of expected outcome

V. Conclusion

In this paper a wide survey has been done on different approaches for privacy preserving data mining. In order to find out better privacy and high data accuracy on high dimensional data randomized noise based perturbation technique with SVD is selected.

Our future work will decide that Privacy preservation and knowledge discovery both goals can be achieved by using these 3 methods.

References

[1] Xingdong Wu and XingQuan Zhu “Mining with Noise Knowledge: Error-Aware Data Mining” IEEE Transaction on System, Man and Cybernatics- Part A: Systems And Humans, Vol. 38 July 2008.

[2] Xindong Wu and Xingquan Zhu “ Mining With Noise Knowledge: Error-Aware Data Mining.” IEEE Transactions On Systems, man, and Cybernatics: JULY 2008.

[3] Xiaolin Zhang, Hongjing Bi “Research on Privacy Preserving classification , data mining based on random perturbation.” IEEE international conference on Information Networking and automation 2010.

[4] Kaniska Bhaduri, Member IEEE, Mark d. Stefanski, and Ashok N. Shrivastava “Privacy-Preserving Outlier Detection through Random Nonlinear Data Distortion.” IEEE Transaction on System Man and Cybernetics Vol. 41, Issue 1, Feb 2011.

[5] Keng-Pie Lin and Ming-Syan Chen “On the Design and Analysis of the Privacy-Preserving SVM classifier.” IEEE Transaction on Knowledge And Data Engineering Vol.23, Issue 11, November 2011.

[6] Guang Li and Yadong Wang “Privacy preserving data mining based on sample selection and singular value decomposition.” IEEE international conference on Internet computing and information services 2011.

[7] Guang Li and Yadong Wang “Privacy preserving classification Method Based on singular value decomposition.” International Arab Journal of information Technology Vol.9, Issue 6, 2012.

[8] Yaping Li, MinghuaChen , Qiwei Li, and Wei Zhang “Enabling Multilevel Trust in Privacy Preserving Data Mining .” IEEE Transactions on Knowledge And Data Engineering, Vol. 24 , Issue-9, September 2012.

[9] Pui k Fong and Jens H. Webber-Jahnke “Privacy Preserving Decision Tree Learning Using Unrealized Data Sets” IEEE Transactions on Knowledge and Data Engineering Vol.24, Issue-2, February 2012.

[10] Priyank jain , pratibha Tapashetti , Dr. A. S.Umesh, sweta sharma: “Privacy preserving processing of high dimensional data classification based on sample selection and SVD.” IEEE international conference on control, Automation, robotics and embedded system 2013.

[11] Alxandre Evfimievski “Randomization in Privacy Preserving Data Mining” SIGKDD Explorations VOL.4 Issue-2 Pg-43-48.

[12] Charu C. Aggarwal “On the Analytical Properties of High Dimensional Randomization.” IEEE Transaction on Knowledge And Data engineering Vol. 25, Issue-7 July 2013.

[13] Mohnish Patel, Prashant Richarya, Anurag Shrivastava "Privacy preserving Using Randomization And Encryption Methods." Scholars journal of Engineering And technology(SJET) 2013, Issue-3 pg.117-121.

[14] Tiancheng Li, Ninghui li "Slicing: A new approach for privacy preserving data publishing." IEEE Transactions on Knowledge and Data Engineering Vol.24, Issue-3, March 2013.

[15] Nivetha.P.R, Thamarai selvi.K "A Survey on Privacy preserving Data Mining Techniques."International Journal of Computer Science and Mobile Computing Vol. 2 Issue-10, October 2013, pg.166-170.

[16] Wenjun Lu , Avinash L. Varna(Member IEEE) & Min Wu "Confidentiality-preserving image Search : A Comparative Study Between Homomorphic Encryption and Distance-preserving Randomization." IEEE Vol-2 2014 pg. 125-141.

[17] Santosh Kumar Bhandare "Data transformation and encryption based privacy preserving Data mining System." International Journal of Advanced Research in Computer Science & Software Engineering Vol. 4, Issue 7, July 2014.

[18] Dhivakar k, Mohana "A Survey on privacy preservation approaches and techniques." International Journal of Innovative Research in Computer & Communication Engineering Vol. 2, Issue 11, November 2014.

[19] Sachin janbandhu, Dr. S.M. Chaware "Survey on Data Mining with privacy preservation" International Journal on Computer Science & Information Technology Vol-5(4) 2014.

[20] S.Sasikala, S. Nathira banu "Privacy preserving data mining using Piecewise Vector Quantization (PVQ)." International journal of Advanced Research in Computer Science and Technology 2014.

[21] Xiaohua Tong, Zhen Ye, Yusheng Xu , ShiJie Liu , LIngyun Li, Huan Xie, and Tianpeng Li "A Novel Subpixel Phase Correlation Method Using Singular Value Decomposition Method and Unified Random Sample Consensus." IEEE Transactions on Geosciences and Remote Sensing Vol.53 Issue-8 August 2015.