Privacy Preserving Data Mining using Concept of Dimensionality Reduction

Gargi Shah [1]

[1] *Assistant Professor,* D*epartment of Computer Engineering, Vadodara Institute of Engineering & Research, Gujarat, India*

ABSTRACT

Privacy-preserving data mining is basically the branch of study of various valid mining models. Privacy-preserving data mining hide sensitive information of the data. If the data is flowing continuously, it would be necessary to rescan the database, which leads to more computation time and unable to respond properly to the user. As per study it is said that accuracy of data is inversely proportional to data transformation. So there is need to develop system which preserves privacy along with accuracy.

Keyword: - *accuracy, privacy, dimensionality reduction, eigen values*.

1. INTRODUCTION

Data mining efficiently discover valuable, non-obvious information from large datasets, is particularly vulnerable to abuse. A fruitful future research leadership in data mining is the development of technology that incorporates the concern for privacy. A recent survey of web users 17% of respondents as privacy fundamentalists, the unclassified data on a site, even if privacy measures are in place [1]. A more recent study of web users found that 86% of respondents believe that information for participation in benefits programs is a matter of individual choice privacy[2].Nowadays organisms around the world are dependent on mining gigantic datasets. These datasets typically contain delicate individual information inevitably all is exposed to the various parties. Consequently privacy issues are constantly in the limelight and the public dissatisfaction may well threaten the exercise of data mining. It is of great importance used technical security to protect the confidentiality of individual values for data mining for the development of appropriate Malthus. There is much research on privacy-preserving data mining (PPDM) [6] malfunctioning, randomization and secure multi-party system based calculations.

Many government agencies, businesses and non-profit organizations to support their short-and long-term schedule activities, to collect for a way to store, analyze and report data on persons, households or businesses looking. Information systems therefore contain confidential information such as social security numbers, income, credit ratings, type of illness, customer purchases, etc., that need to be adequately protected. With the Web revolution and the emergence of data mining, have privacy concerns provided technical challenges fundamentally different from those that occurred before the information age [3].

One of the sources of privacy violation is called data magnets. Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the user's control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

Securing against unauthorized access has been a long-term goal of the database security research community and the government research statistical agencies. Solutions to such a problem require combining several techniques and mechanisms. In an environment where data have different sensitivity levels, this data may be classified at different levels, and made available only to those subjects with an appropriate clearance.

1.1 CLUSTERING

Clustering is a well-known problem in statistics and engineering, namely, how to arrange a set of vectors (measurements) into a number of groups (clusters). Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization. The problem has been formulated in various ways in the machine learning, pattern recognition optimization and statistics literature. The fundamental clustering problem is that of grouping together (clustering) data items that are similar to each other. Given a set of data items, clustering algorithms group similar items together. Clustering has many applications, such as customer behavior analysis, targeted marketing, forensics, and bioinformatics [7].

Small companies have recognized the value in data, especially with the introduction of the knowledge discovery process. However, small companies do not have enough expertise for doing data analysis, although they have good domain knowledge and understand their data.

2. DIMENTIONALITY REDUCTION

Let's say you are measuring three things: age, hours on internet and hours on mobile. There are 3 variables so it is a 3D data set. 3 dimensions are an x, y and z graph, It measure width, depth and height (like the dimensions in the real world). Now imagine that the data forms into an oval like the ones above, but that this oval is on a plane. i.e. all the data points lie on a piece of paper within this 3D graph (having width and depth, but no height).

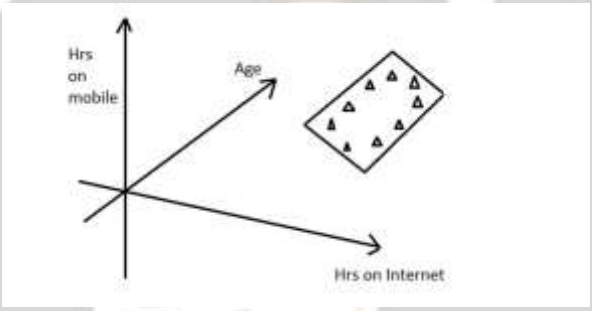

Figure 1: 3D graph

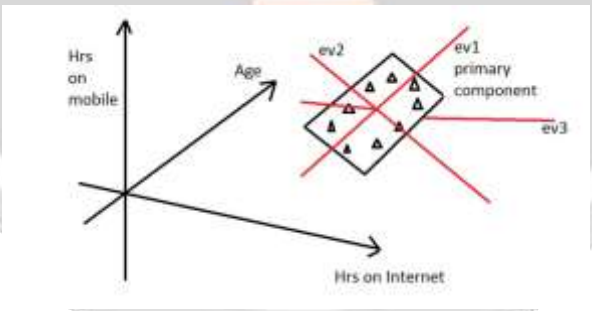

Figure 2: 3D graph with eigen vectors

Like this when we find the 3 eigenvectors/values of the data set (remember 3D problem = 3 eigenvectors), 2 of the eigenvectors will have large eigen values, and one of the eigenvectors will have an eigen value of zero. The first two eigenvectors will show the width and depth of the data, but because there is no height on the data (it is on a piece of paper) the third eigen value will be zero. On the picture below ev1 is the first eigen vector (the one with the biggest eigen value, the principal component), ev2 is the second eigenvector (which has a non-zero eigen value) and ev3 is the third eigenvector, which has an eigen value of zero.

We can now rearrange our axes to be along the eigenvectors, rather than age, hours on internet and hours on mobile. However we know that the ev3, the third eigenvector, is pretty useless. Therefore instead of representing the data in 3 dimensions, we can get rid of the useless direction and only represent it in 2 dimension.

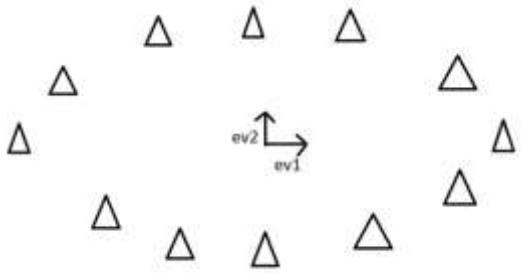

Figure 3: Dimension reduction from 3D to 2D

Note that we can reduce dimensions even if there isn't a zero eigen value. Imagine we did the example again, except instead of the oval being on a 2D plane, it had a tiny amount of height to it. There would still be 3 eigenvectors, however this time all the eigen values would not be zero. The values would be something like 10, 8 and 0.1. The eigenvectors corresponding to 10 and 8 are the dimensions where there is a lot of information; the eigenvector corresponding to 0.1 will not have much information at all, so we can therefore discard the third eigenvector again in order to make the data set simpler.

3. CONCLUSION

In this paper the approach of dimensionality reduction perturbation privacy preserving of group data is presented. It reduces data from 3D to a 2D problem, getting rid of a dimension. Reducing dimensions helps to simplify the data and makes it easier to visualize. It preserve the important geometric properties, thus most data mining models that search for geometric class boundaries are well preserved with the perturbed data. It transforms data in such a way that introduces new challenges in evaluating the privacy guarantee for multidimensional perturbation. It is observed that accuracy is maintained along with privacy.

4. REFERENCES

[1] Majid Bashir Malik and M. Asger Ghazi and Rashid Ali ;"*Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects*"; Third International Conference on Computer and Communication Technology; 978-0-7695-4872-2/12 $26.00 © 2012 IEEE

[2] Hitesh Chhinkaniwala and Dr. Sanjay Garg "*Privacy Preserving Data Mining Techniques: Challenges & Issues*" in Proceedings of International Conference on Computer Science & Information Technology, CSIT – 2011,p.609

[3] Chirag N. Modi,Udai Pratap Rao and Dhiren R.Patel "*Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining",* in 2010 Second International conference on Computing, Communication and Networking Technologies

[4] W.T. Chembian1, Dr. J.Janet, "*A Survey on Privacy Preserving Data Mining Approaches and Techniques*",in Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010,6 February 2010, Chennai, India

[5] Xiaolin Zhang and Hongjing Bi; "*Research on Privacy Preserving Classification Data Mining Based on Random Perturbation*"; International Conference on Information, Networking and Automation (ICINA); 978-1-4244-8106-4/$26.00 © 2010 IEEE

[6] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun ;"*Privacy-Preserving Classification of Data Streams*"; Tamkang Journal of Science and Engineering; Vol. 12, No. 3, pp. 321_330 (2009)

[7] M. Naga lakshmi, K Sandhya Rani;" *Privacy Preserving Clustering Based on Discrete Cosine Transformation*" International Journal of Innovative Research in Science, Engineering and Technology; ISSN: 2319-8753 Vol. 2, Issue 9, September 2013

[8] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, and Thomas Seidl "*MOA: Massive Online Analysis, a Frameworkfor Stream Classification and Clustering.*"

[9] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino,Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis "*State-of-the-art in Privacy Preserving Data Mining*"

[10] Haisheng Li;" *Study of Privacy Preserving Data Mining*" ;Third International Symposium on Intelligent Information Technology and Security Informatics; 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[11] Jian Wang ,Yongcheng Luo ,Yan Zhao Jiajin Le;" *A Survey on Privacy Preserving Data Mining*";2009 First International Workshop on Database Technology and Applications; 978-0-7695-3604-0/09 $25.00 © 2009 IEEE

[12] MohammadReza Keyvanpour, Somayyeh Seifi Moradi;" *Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework*";International Journal on Computer Science and Engineering (IJCSE); ISSN : 0975-3397 Vol. 3 No. 2 Feb 2011

[13] Keke Chen Ling Liu*," A Random Rotation Perturbation Approach to Privacy Preserving Data Classification*"

[14] S. Kasthuri, T. Meyyappan;" *Detection of Sensitive Items in Market Basket Database using Association Rule Mining for Privacy Preserving*";Proceedings of the 2013 International Conference on PRIM, February 21-22; 978-1-4673-5845-3/13/$31.00©2013 IEEE

[15] Nikunj H. Domadiya;" *Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database*";978-1-4673-4529-3/12/$31.00_c 2012 IEEE

[16] Rahena Akhter, Rownak Jahan Chowdhury, Keita Emura, Tamzida Islam, Mohammad Shahriar Rahman, Nusrat Rubaiyat;" *Privacy-Preserving Two-Party k-Means Clustering in Malicious Model*";2013 IEEE 37th Annual Computer Software and Applications Conference Workshops; 978-0-7695-4987-3/13 $26.00 © 2013 IEEE

[17] Jaideep Vaidya , Basit Shafiq ;" *A Random Decision Tree Framework for Privacy-preserving Data Mining*";1545-5971/13/$31.00 © 2013 IEEE

[18] Hitesh Chhinkaniwala and Sanjay Garg ;" *Tuple Value Based Multiplicative Data Perturbation Approach To Preserve Privacy In Data Stream Mining*";International Journal of Data Mining & Knowledge Management Process (IJDKP); DOI : 10.5121/ijdkp.2013.3305; Vol.3, No.3, May 2013

[19] Data Mining concepts and Techniques by Jiawei Han, Micheline Kamber –Elsevier.

[20] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Bank.

[21] www.ittc.ku.edu/~nivisid/**WEKA_MANUAL**.pdf

[22] http://www.cs.waikato.ac.nz/ml/weka/documentation.html

[23] www.cs.waikato.ac.nz/~abifet/**MOA**/**Manual**.pdf

[24] http://moa.cms.waikato.ac.nz/

[25] http://en.wikipedia.org/wiki/Perturbation_theory

[26] https://netbeans.org/features/index.html