ISSN No. 0976-5697

Volume 4, No. 3, March 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Privacy Preserving Data Mining through New Proposed "Two Phase SVD Matrix Factorization Model"

Priyank Jain
Dept. of Information Technology
RGTU University TIT Bhopal Anand
Nagar Bhopal MP India
Priyankjain1984@gmail.com

Bhupesh Gour, Asif Ullah Khan
Dept. of Computer Sc. & Engineering
RGTU University TIT Bhopal
Anand Nagar
Bhopal MP India

Abstract- In recent years, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the Internet. In particular,recent advances in the data-mining field have lead to increased concerns about privacy. While the topic of privacy has been traditionally studied in the context of cryptography and information hiding, recent emphasis on data mining has lead to renewed interest in the field. A number of algorithmic techniques have been designed for privacy-preserving data mining. We have proposed new model Two-Phase SVD Matrix Factorization Model, which provides different level of privacy.

Keywords- Privacy preserving, Data mining, svd, matrix factorization.

I. INTRODUCTION

Maintaining data mining accuracy is an important issue in privacy preserving data mining. Using Matrix Factorization method, we propose new approach Two- Phase Matrix Factorization Model, Matrix Factorization Model have following advantage:

a. Better Performance than other approaches Matrix factorization technique Singular Value Decomposition (SVD)is considered to be promising technique for privacy preserving data mining.
b. Matrix factorization achieves both high level privacy preservation and high degree data mining accuracy.
c. It has also high level addressing Accuracy.

Previous work [2] on two-phase perturbation that allows each individual to choose his/her own privacy level according to his or her privacy choices. I have proposed new approach "Two Phase SVD Matrix Factorization Model", which is based on SVD Matrix Factorization Method & provide different level of privacy.This model has better performance & better accuracy than other Model

II. RELETED WORK

Jaideep Vaidya Chris Clifton Work describe "Ever-increasing data collection, along with the influx of analysis tools capable of handling huge volumes of information, has led to privacy concerns[1]. Jie Wang and Jun Zhang work describe" With better performance than some classical data perturbation approaches nonnegative matrix factorization and singular value decomposition are considered to be promising techniques for privacy preserving data mining. Experimental results demonstrate that mining accuracy on the distorted data used these methods is almost as good as that on the original data, with added property of privacy preservation. It indicates that the matrix factorization-based data distortion scheme perturb only confidential attributes to meet privacy requirements while preserving general data pattern for knowledge extraction"[3]. Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham work describe this work, propose an individually adaptable perturbation model, which enables the individuals to choose their own privacy level. Hence this model provides different privacy guarantees for different privacy preferences"[2].

III. TWO PHASE SVD MATRIX FACTORIZATION MODEL

We can describe Two-Phase Matrix Factorization Model as follows:

a. The system first breakdown original data into factorize data using Singular Value Decomposition (SVD) Matrix Factorization Method. The Original data Factorize into factorize data [U] n*n (left singular vector), [V*] m*m (Right singular vector) & Factorize Data [Σ] n*m.
b. User i chooses his/her privacy level among various privacy levels.
c. Based on the user's privacy level choice, the system applies an interval length that correspondent to the chosen privacy level. Later on, W' is created by sampling uniformly from the interval
d. W' value is sent to the data miner.

This model is Two-Phase Matrix Factorization model, the means of Two Phase i.e. not only apply the matrix factorization method to data but also provide the desire level of privacy: High, Low. To apply the matrix factorization, I have used Singular Value Decomposition Method.

CONFERENCE PAPER II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

A. Singular Value Decomposition (SVD):

The SVD of the matrix A can be written as [A]n*m =[U] n*n × [Σ] n*m × [V*]m*m U and V contain the left and right singular vectors of A, respectively, and the diagonal of z is the singular values in descending order. These three matrices reflect a breakdown of the original relationship into linearly independent vectors. Using a truncated SVD, [A]n*m =[U] n*n × [Σ] n*m × [V*]m*m the dimensionality of the data can be reduced by projecting the m column vectors onto a k dimensional space corresponding to the k largest singular values. Therefore, it is possible to achieve higher-level data mining accuracy by performing the truncated SVD operation on the original data.

We have applied Singular Value Decomposition (SVD)[5] because it is able to achieve higher-level accuracy. the singular value decomposition (SVD) is an important factorization of a rectangular real or complexmatrix, with several applications in signal processing and statistics. Applications which employ the SVD include computing the pseudoinverse, least squares fitting of data matrix approximation, and determining the rank of a matrix.

B. Different level of Privacy:

There are two privacy Level in Two Phase Matrix Factorization Model.

a. High Level Privacy SVD Matrix Factorization Model:

High Level Privacy Matrix Factorization Model concerned with high-level security & also maintains high accuracy. First we apply data to SVD, The Original data Factorize into factorize data [U] n*n (left singular vector), [V*] m*m (Right singular vector) & Factorize Data [Σ] n*m.This Matrix contains p records of same dataset. Each record has q attributes. The Zero Value Shows null Value.

Now Matrix is factorized:

[A]n*m= [U] n*n× [Σ] n*m × [V*] m*m

In Matrix U & V* Apply the Noisy distribution individually by Matrix Addition.

Noisy distribution: The random variable has a normal distribution with $\mu = 0$

U' = U + l1

Where U= left singular vector Matrix, U'= left Matrix after apply the Noisy Distribution.

V*'= V* + l2

Where V*= Right singular vector Matrix, V*'= Right Matrix after apply the Noisy Distribution, l2 = Noisy Distribution Matrix

Now we get the value U', V*' & Σ. We have applied Noisy Distribution to dataset to get High Privacy compare to other methods & it also provides the High level of Security. The Value U', V*' & Σ sent to server. To get Original value of U & V* at Server apply Agrawal-Bayes[7][8] estimation Algorithm.

$$f_x^j(a) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z)\mathrm{d}z}$$

Initial f_X^0 = Uniform distribution, Iteration number j: = 0

$$f_x^{j+1}(a) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z)\mathrm{d}z}$$

j = j + 1;

After Getting Original value of U & V*, Matrix multiplication take place here Between U, Σ & V* & we get the Original value of dataset.

b. Low Level Privacy SVD Matrix factorization Model:

In Low Level Privacy Matrix Factorization Model, only apply the Singular Value Decomposition (SVD) to dataset, it provides low level of privacy & High accuracy

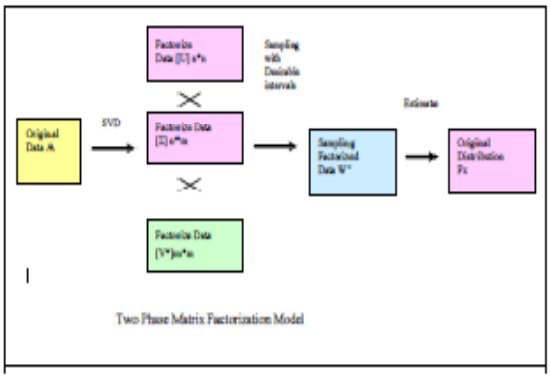

Fig 1. Two Phase Matrix Factorization Model

IV. IMPLEMENTATION & EXPERIMENT RESULT

Two-Phase SVD Matrix factorization Model in Privacy Preserving Data Mining has been implemented using Java Technology. This is based on Client Server Application. Both Client and Server have minimum 1.7 GHz. Pentium IV machine with 256 MB RAM running windows operation system & version of Java JDK1.6 installed in both Client & Server Machine. To establish connection between Client & Server use the port number. To perform this work, I used Vehicle dataset. Detail description of Vehicle dataset is given below.

A. Data Set Description:

The goal of this work is not only achieve, High-level privacy as well as Low-level privacy, for this purpose taken the Vehicle Dataset. It is numeric dataset. The source of dataset is Turing Institute, Glasgow, Scotland. JP Siebert originally gathered this data at the TI in 1986-87. The original purpose was to find a method of distinguishing 3D objects within a 2D image. Vehicle Dataset contain around 52000 samples, after training taken 10000 samples. It has 5 attributes Compactness, Circularity, Distance Circularity, Radius Ratio, Axis Aspect Ratio.

B. Experiment Result:

In experiment we apply different size of dataset like 2KB, 8KB, 16KB, 32KB &

Percentage of accuracy correspondences we obtain. Figure 2 showing the Comparison between High Level Privacy SVD Matrix Factorization Model & Low Level Privacy SVD Matrix Factorization Model.

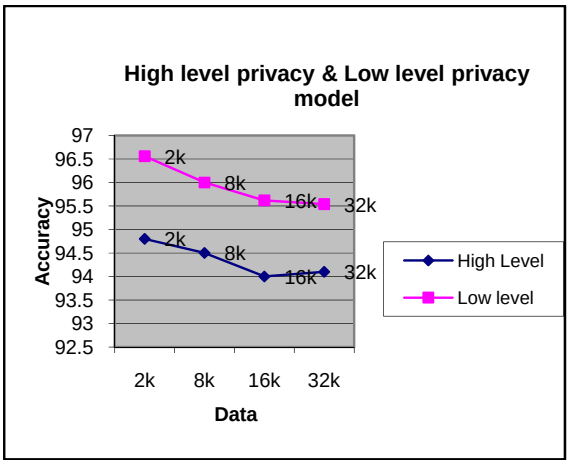

Figure 2. Comparison between High Level Privacy SVD Matrix Factorization Model & Low Level Privacy SVD Matrix Factorization Model.

Low level of privacy perform significantly better than the High Level of Privacy in Two phase SVD matrix factorization model according to figure 2 in our experiment. High level Privacy SVD Matrix factorization obtain average 94.22% Accuracy & Low level Privacy SVD Matrix factorization obtain average 96.1% Accuracy according to our experiment.

V. CONCLUSION & FUTURE WORK

Privacy-Preserving Data Mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information.

I proposed new approach Two-phase matrix Factorization model, which Provide High Accuracy & High, Low privacy according to user requirement. Low level of privacy perform significantly better than the High Level of Privacy in Two phase SVD matrix factorization model in our experiment work. Future work is concerned for studying different types of Privacy Preserving Data Mining Methods & generating different levels of Privacy models because one privacy level is not the need of modern society.

VI. REFERENCES

[1]. Jaideep Vaidya, Chris Clifton: Privacy-Preserving Data Mining: Why, How, and When. IEEE Security & Privacy 2(6): 19-27 (2004).

[2]. The applicability of the perturbation based privacy preserving data mining for real- world data. IEEE, Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham 2007.

[3]. R. Agrawal, R. Srikant, Privacy-preserving data mining, in: SIGMOD Conference,2000, pp. 439–450.

[4]. Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. IEEE Symp. on Security and Privacy, 1998.

[5]. Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization.IEEE, Jie Wang and Jun Zhang 2007.

[6]. C. J. Lin, "Projected gradient methods for non-negative matrix factorization, "http://www.csie.ntu.edu.tw/Hcjlin/papers/pgradnmf.pdf

[7]. Machanavajjhala A., Gehrke J., Kifer D. l-diversity: Privacy beyond k-anonymity. IEEE ICDE Conference, 2006.

[8]. Rizvi S., Haritsa J.Maintaining Data Privacy in Association Rule Mining. VLDB Conference, 2002.