IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Privacy Preserving in High Dimensional Data using Randomized and SVD algorithm

Tripti Singh Thakur[1], Dr. Abha choubey[2]

PG Scholar, Department of CSE, SSGI, SSTC, Bhilai, C.G

Associate Professor, Department of CSE, (FET), SSGI, SSTC, Bhilai, C.G

Abstract: Data mining is a process of extracting useful knowledge and hidden confidential patterns from data. The growth of information technology increases need for electronic data to be carefully stored and secretly shared. In last few decades a wide variety of approaches and techniques have been proposed for modifying data in such a way that privacy will remain preserved. Perturbation based randomization and SVD are one of the methods for preserving privacy. We need an algorithm which protects sensitive private information from huge databases in data mining system. In this paper we proposed an efficient privacy preserving data mining technique using perturbation based randomization in combination with SVD. In this technique we will apply several classification schemes on perturbed data. Experimental comparisons will define the effectiveness of this algorithm.

Keywords: Randomization; perturbation; SVD.

I. INTRODUCTION

Data mining system having large amount of private and confidential data, which needs to be preserved in such a way that data confidentiality remains in data. For the process of picking out information for databases some techniques like- clustering, classification, associations are used. Data privacy can be preserved in two ways- By protecting the Data, by protecting the result of the data. Its main goal is to develop efficient algorithms that allow one to extract relevant knowledge from large amount of data, while sensitive information will be prevented from inferences. The concept of privacy preserving data mining preserve Personal information from different data mining algorithms in such a way that the private data and private knowledge remain confidential after the mining process.

Data cleansing effects only on certain type of errors and results imperfect data, noisy data elimination leads to some information loss, So an Error aware data mining design is considered which takes advantage of statistical error information to improve data mining result [1].Non linear data distortion using potentially nonlinear random data transformation is used for anomaly detection from sensitive datasets [3]. SVM classifier is analyzed for privacy violation problem and post processed SVM is applied [4].SVD based perturbation method and sample selection is also used in combination for preserving privacy [6]. Perturbation based PPDM is also expanded to multilevel trust for preventing diversity attacks [7]. Decision tree learning using unrealized data sets is also useful in preserving privacy [8]. Privacy can also be preserved by decision tree learning on unrealized data sets [9]. Analytical properties of high dimensional randomization have also been studied for examining strengths and weaknesses of randomization [11]. Data privacy is also preserved by using randomization with encryption method [12]. Slicing preserve better data utility and can be used for membership disclosure protection [13]. SVD based data transformation methods have also been used for privacy preserving clustering [14]. Distance preserving randomization is used in combination with homomorphic Encryption technique to preserve confidentiality in image search [16].

Data perturbation combined with cryptographic technique also provides efficient results[17].If data transformation and encryption techniques are applied in combination then the data privacy is preserved strongly [18].In order to deal with privacy preserving clustering piece vise quantization approach is used here after encoding original data cannot be revealed hence privacy is preserved[20]. A novel sub pixel phase correlation method is used in combination with SVD and unified random sample consensus (RANSAC) for preserving privacy in applications like remote sensing community [21].

II. PROBLEM DEFINITION

Data privacy has been an active research topic for the last few decades because sensitive information increases in database. Sometimes due to law (for medical databases) privacy is needed or sometimes it can be influenced by interest. Data mining field, connecting the three worlds of databases, Artificial Intelligence and statistics so it is important to develop such an algorithm for privacy preservation that the data accuracy and integrity remains in data without affecting the data confidentiality.

III. METHODOLOGY

Main idea behind privacy preserving data mining is to develop such solution which will result data security with data consistency and confidentiality with low

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

computational complexity. In this paper we proposed a method which uses perturbation based randomization method combined with SVD. In perturbation based randomization technique the original data is modified by randomly adding some values i.e. noise to original data which is independent of the behavior of other records. It provides some deeper statistical approach to security and privacy. Perturbation based randomization technique perturb data element or attribute randomly by addition, by multiplication or by combination of both to original datasets, the randomly modified data is referred as noisy data. Perturbation based randomization technique is Easy to implement and provide High search accuracy. It is computationally efficient and Suitable for different user requirements. SVD technique is a matrix factorization technique which perturbs every sample of data to the same degree and Provide less information loss.

IV. DATA SET DESCRIPTION

Data has been taken from University of California (UCI), Machine learning repository. Datasets are Glass, Iris and wine datasets. The performance measure of original Glass Data set is given below.

Table 1 Performance Measure for original Glass data in percentage

Classifier	Performance measure for original glass data				
	accuracy	precision	recall	f_meas-ure	gmean
Ensemble	90.74	76.47	92.85	83.87	91.41
Naïve Bayes	87.03	76.47	81.25	78.78	85.26
KNN	88.88	75	85.71	80	87.83
SVD	90.71	72.22	1	83.87.	93.07

V. RESULT

We have implemented an algorithm using Perturbation based Randomized technique in combination with SVD. For every training sample set, we used four classifier. Firstly the Original dataset is passed to all four classifiers after applying SVD and secondly the Noisy data is passed to all four classifiers after SVD. The four classifiers are Ensemble, NaiveBayes, KNN and SVM classifier. Table 1 and Table 2 shows the performance measures after Evaluation of the all these methods in original data and Noisy data respectively.

Table 2 Performance measures after Evaluation of Original Data in percentage

Classifier	**Performance Measures for Original Glass data With SVD**				
	Accuracy	**Precision**	**Recall**	**Fmeas ure**	**G mean**
Ensemble	92.59	85.71	85.71	85.71	90.2
Naïve Bayes	88.8	70.5	92.3	80	90
KNN	92.59	85.71	85.71	85.71	90.2
SVM	100	2	2	2	2

Table 3 Performance measures after Evaluation of Noisy Data in percentage

Classifier	**Performance Measure of Noisy Glass data With SVD**				
	Accuracy	**Precision**	**Recall**	**Fmeas ure**	**G mean**
Ensemble	94.44	92.30	85.71	88.88	91.4
Naïve Bayes	94.44	84.61	91.66	88	93.43
KNN	92.59	85.71	85.71	85.71	90.23
SVM	100	2	2	2	2

The performance measures in table 2 and table 3 are Accuracy, Precision, Recall, F_measure and Gmean.
Accuracy is defined as the closeness of a measurement to the true value. We can say that accuracy is proportion of true sample among total number of samples examined

FP rate = False Positive / N; N is the number of negative samples
TP rate = True Positive / P; P is the number of positive samples
Accuracy = (TP+ FP) / (P + N);

Figure 1 shows the accuracies of original and noisy data trained on classifiers.

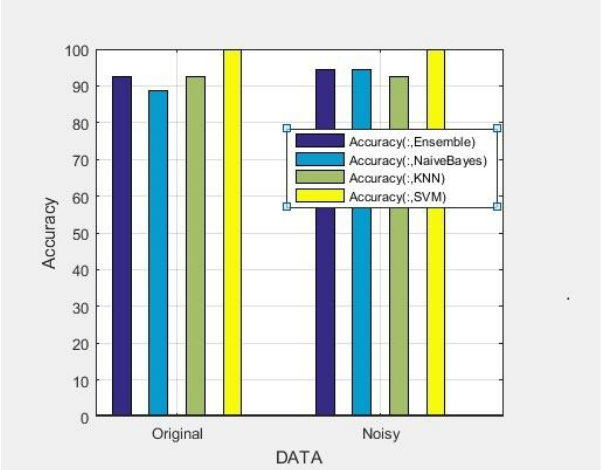

Figure 1 accuracies of original and noisy data trained on classifiers

Precision is also known as positive predictive value it defines the probability that a randomly selected document is relevant.

Precision =TP / (TP+ FP);

Figure 2 shows the Precision of original and noisy data trained on classifiers.

Recall is also known as sensitivity it is proportion of the sample that tested positive and are positive of all the sample that actually are positive means it defines that a randomly selected document is retrieved in a search.

Recall = TP / P;

DOI 10.17148/IJARCCE.2016.55213

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

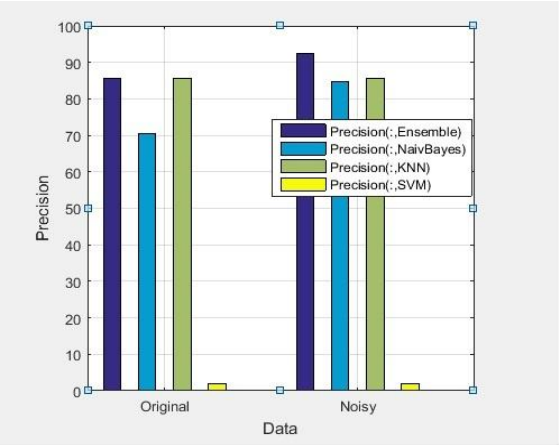

Figure 2 Precision of original and noisy data trained on classifiers

Figure 3 shows the recall of original and noisy data trained on classifiers.

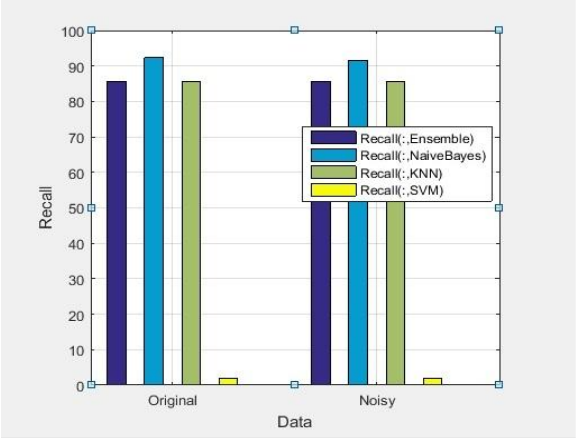

Figure 3 Recall of original and noisy data trained on classifiers

F_measure is a measure that combines precision and recall i.e. it is harmonic mean of precision and recall.
F_measure = Precision * Recall;

Figure 4 shows the F_measure of original and noisy data trained on classifiers.

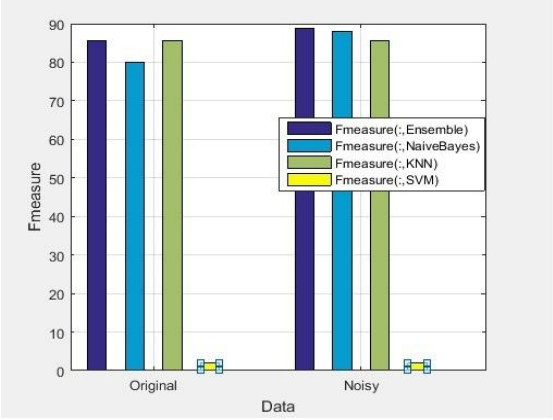

Figure 4 F_measure of original and noisy data trained on classifiers

Gmean predicts for a classifier i.e. the value of Gmean predicts for a good classifier.

Gmean = square root of(TP/P*TN/N);

Figure 5 shows the Gmean of original and noisy data trained on classifiers.

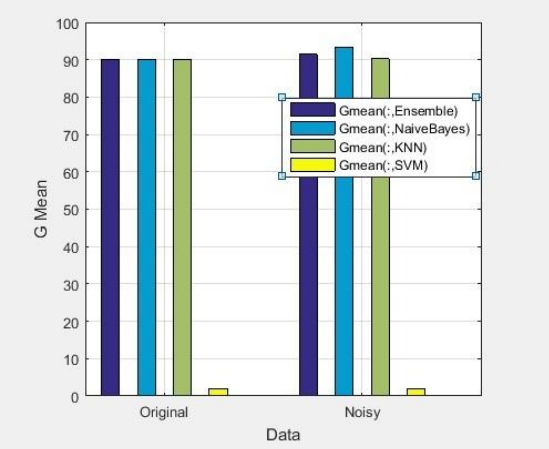

Figure 5 G_mean of original and noisy data trained on classifiers

VI. CONCLUSION

We have taken a look on Perturbation based randomized technique in combination with SVD. Different classifier results different accuracy and recall values for the same data set. Singular value Decomposition has big role in privacy preserving data mining classification. It results high search accuracy with less computational complexity and information loss. In summery we can say that perturbation based randomization in combination with SVD takes a significant step towards making privacy preserving data mining classification. In future some other decomposition methods can be applied for preserving privacy in data mining.

REFRENCES

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.
[8] 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.
[9] 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.
[10] Alxandre Evfimievski "Randomization in Privacy Preserving Data Mining" SIGKDD Explorations VOL.4 Issue-2 Pg-43-48.
[11] Charu C. Aggarwal "On the Analytical Properties of High Dimensional Randomization." IEEE Transaction on Knowledge And Data engineering Vol. 25, Issue-7 July 2013.
[12] 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.
[13] 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.
[14] M. Naga Laxmi & K. Sandhya rani "SVD based Data Transformation Methods for privacy preserving clustering."International journal of computer applications vol- 78, no-3, September 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.