Homeland Defense, Privacy-Sensitive Data Mining, and Random Value Distortion

Souptik Datta[*] Hillol Kargupta[†] Krishnamoorthy Sivakumar[‡]

Abstract

Data mining is playing an increasingly important role in sifting through large amount of data for homeland defense applications. However, we must pay attention to the privacy issues while mining the data. This has resulted in the development of several privacy-preserving data mining techniques. The random value distortion technique is one among them. It attempts to hide the sensitive data by randomly modifying the values. This paper questions the utility of the random value distortion technique. The paper develops a random matrix-based spectral filtering technique to retrieve original data from the dataset distorted by adding random values. The proposed method works by comparing the spectrum generated from the observed data with that of random matrices. The paper presents the theoretical foundation and extensive experimental results to demonstrate that the random value distortion technique may not preserve any data privacy after all.

Keywords: Privacy preserving data mining, value perturbation, random matrices, eigenanalysis.

1 Introduction

Many homeland defense applications require mining heterogeneous data for creating profiles, constructing social network models, detecting terrorist communications among others. Usually the data is very sensitive to privacy issues. Financial transactions, healthcare records, and network communication traffic are a few examples. Data mining in such privacy-sensitive domains is facing growing concerns. Therefore, we need to develop data mining techniques that are sensitive to the privacy issue. This has fostered the development of a class of data mining algorithms [4, 13] that try to protect the data privacy with varying degrees of success. Most of these algorithms try to extract the data patterns without directly accessing the original data and guarantees that the mining process does not get sufficient information to reconstruct the original data.

This paper explores the random value perturbation-based approach [4], a well-known technique for masking the data using random noise. The idea is to preserve data privacy by adding random noise, while making sure that the random noise still preserves the signal from the data so that the patterns can be closely estimated. This paper questions the privacy-preserving capability of the random value perturbation-based approach and shows that the original data can be accurately estimated from the perturbed data using a spectral filter that exploits some theoretical properties of random matrices. It presents the theoretical foundation and provides experimental results to support this claim.

Section 2 offers an overview of the related literature in privacy preserving data mining. Section 3 describes the random data perturbation method proposed in [4]. Section 4 discusses the theoretical foundation of our approach that relies on known properties of random matrices. Section 5 describes the random matrix-based eigen analysis methods to extract the original dataset. Section 6 applies the proposed technique and reports its performance for various data sets. Finally, Section 7 concludes this paper and outlines future research directions.

2 Related Work

Privacy is related to an individual or groups of individuals sharing some common features in a given context. Preserving privacy of the data is important in data mining applications dealing with sensitive data. Internet marketing firms, that have access to sensitive financial data, retain their ability to analyze the data with a considerable emphasis on privacy preservation [2]. A “privacy policy” is almost a recognized standard for operating with private information for such companies. There are different techniques for this. For example, if the privacy is associated with the identity of an individual then sometimes removing the identification in-

[*] Computer Science and Electrical Engineering Department, University of Maryland Baltimore County, USA.

[†] Computer Science and Electrical Engineering Department, University of Maryland Baltimore County, USA.

[‡] School of Electrical Engineering and Computer Science, Washington State University, USA.

formation from the data solves the problem. However, there exist many applications where such simple solutions do not work. The data set may still reveal certain information that violates the privacy of different entities associated with the data. Therefore, data mining techniques are necessary that can work without directly accessing the raw data.

There exists a growing body of literature on this topic. Cryptographic tools are suggested in [21] in order to secure data transmission, along with communication between local sites as opposed to one centralized site. A privacy preserving technique to construct decision trees [17] is reported in [14]. The approach depends on a completely reliable intermediary party, in order to regulate the privacy preservation. Kantarcioglu and Clifton [13] investigate an association rule mining from homogeneous data using a commutative encryption tool.

Two very general procedures are frequently executed in privacy preserving tasks are query restriction limitation, and data perturbation. Query restriction limits the amount of information released based on the amount of information available. On the other hand, data perturbation is conducted in a manner to add noise to the original data so that its actual information cannot be extracted. A value distortion based technique fro data perturbation is suggested in [4]. Adding noise to the values of a database is reported in [20].

Several related works concerning mining association rules also have similar privacy preserving aims. The idea of sensitive rules, as well as privatized patterns are introduced in [1, 23, 3, 7]. Furthermore a group of related literature exists, which, focus not on preserving the original data, but rather the underlying patterns in the data such as in [11, 6]. Such works consider original data to be non-sensitive information, however the holders of this data does not want clients to infer certain patterns, holding the patterns, not the original data as the sensitive issue. [3] attempts to distort the original information, by removing certain item sets such that particular patterns cannot be detected, with minimal effect on the overall database and alternate item sets.

3 Random Value Perturbation Technique: A Brief Review

For the sake of completeness, we now briefly review the random data perturbation method suggested in [4]. We also discuss the procedure for reconstructing the original data distribution, as suggested in [4].

3.1 Perturbing the Data

The random value perturbation method attempts to preserve privacy of the data by modifying values of the sensitive attributes using a randomized process [4]. The authors explore two possible approaches - Value-Class Membership and Value Distortion - and emphasize the Value Distortion approach. In this approach, the owner of a dataset returns a value $x_i + r$, where x_i is the original data, and r is a random value drawn from a certain distribution. Most commonly used distributions are the uniform distribution over an interval $[-\alpha, \alpha]$ and Gaussian distribution with mean $\mu = 0$ and standard deviation σ. The n original data values $x_1, x_2, \ldots, x_n$ are viewed as realizations of n independent and identically distributed (i.i.d.) random variables X_i, $i = 1, 2, \ldots, n$, each with the same distribution as that of a random variable X. In order to perturb the data, n independent samples $r_1, r_2, \ldots, r_n$, are drawn from a distribution R. The owner of the data provides the perturbed values $x_1 + r_1, x_2 + r_2, \ldots, x_n + r_n$ and the cumulative distribution function $F_R(r)$ of R. The reconstruction problem is to estimate the distribution $F_X(x)$ of the original data, from the perturbed data.

3.2 Estimation of Distribution Function from the Perturbed Dataset

The authors [4] suggest the following method to estimate the distribution $F_X(x)$ of X, given n independent samples $w_i = x_i + r_i$, $i = 1, 2, \ldots, n$ and $F_R(r)$. Using Bayes' rule, the posterior distribution function $F'_X(x)$ of X, given that $X + R = w$, can be written as

$$F'_X(x) = \frac{\int_{-\infty}^{x} f_Y(w-z) f_X(z) dz}{\int_{-\infty}^{\infty} f_Y(w-z) f_X(z) dz},$$

which upon differentiation with respect to x yields the density function

$$f'_X(x) = \frac{f_Y(w-x) f_X(x)}{\int_{-\infty}^{\infty} f_Y(w-z) f_X(z) dz}.$$

If we have n independent samples $x_i + r_i = w_i$, $i = 1, 2, \ldots, n$, the corresponding posterior distribution can be obtained by averaging:

$$f'_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - x) f_X(x)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z) dz}. \qquad (1)$$

For sufficiently large number of samples n, we expect the above density function to be close to the real density function $f_X(x)$. In practice, since the true density $f_X(x)$ is unknown, we need to modify the right-hand side of equation (1). The authors suggest an iterative procedure where at each step $j = 1, 2, \ldots$, the posterior density $f_X^{j-1}(x)$ estimated at step $j-1$ is used in the right-hand side of equation (1). The uniform density is used to initialize the iterations. The iterations are carried out until the difference between successive estimates becomes small. In order to speed up computations, the authors also discuss approximations to the above procedure using partitioning of the domain of data values.

4 Theory of Random Matrices

In this section, we discuss the general theory of random matrices that is used to filter the noise from the perturbed dataset to obtain an estimate of the actual dataset. Our filtering approach is based on the observation that the distribution of eigenvalues of random matrices [16] exhibit some well known characteristics.

A random matrix is a matrix whose elements are random variables with given probability laws. The theory of random matrices deals with the statistical properties of the eigenvalues of such matrices. Eigenvalues of random matrices offer many interesting properties. For example, Wigner's semicircle law, which says if X is an $n \times n$ matrix and has i.i.d. entries with zero mean and unit variance, the distribution of eigenvalues of $\frac{X+X'}{2\sqrt{2n}}$ has a probability density function given by

$$f(x) = \begin{cases} \frac{1}{\pi}(2n - x^2)^{1/2}, & |x| < \sqrt{2n} \\ 0, & \text{otherwise}. \end{cases}$$

In this paper, we are mainly concerned about distribution of eigenvalues of the sample covariance matrix obtained from a random matrix. Let X be a random $m \times n$ matrix whose entries are X_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n$, are i.i.d. random variables with zero mean and variance σ^2. The covariance matrix of X is given by $Y = \frac{1}{m} X'X$. Clearly, Y is an $n \times n$ matrix. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of Y. Let

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} U(x - \lambda_i),$$

be the empirical cumulative distribution function (c.d.f.) of the eigenvalues λ_i, $(1 \leq i \leq n)$, where

$$U(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is the unit step function. In order to consider the asymptotic properties of the c.d.f. $F_n(x)$, we will consider the dimensions $m = m(N)$ and $n = n(N)$ of matrix X to be functions of a variable N. We will consider asymptotics such that in the limit as $N \to \infty$, we have $m(N) \to \infty$, $n(N) \to \infty$, and $\frac{m(N)}{n(N)} \to Q$, where $Q \geq 1$. Under these assumptions, it can be shown that [12] the empirical c.d.f. $F_n(x)$ converges in probability to a continuous distribution function $F_Q(x)$ for every x, whose probability density function (p.d.f.) is given by

$$f_Q(x) = \begin{cases} \frac{Q\sqrt{(x-\lambda_{\min})(\lambda_{\max}-x)}}{2\pi\sigma^2 x} & \lambda_{\min} < x < \lambda_{\max} \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where $\lambda_{\min} = \sigma^2(1 - 1/\sqrt{Q})^2$ and $\lambda_{\max} = \sigma^2(1 + 1/\sqrt{Q})^2$. Further refinements of this result and other discussions can be found in [19, 9, 15, 5, 8, 22, 18].

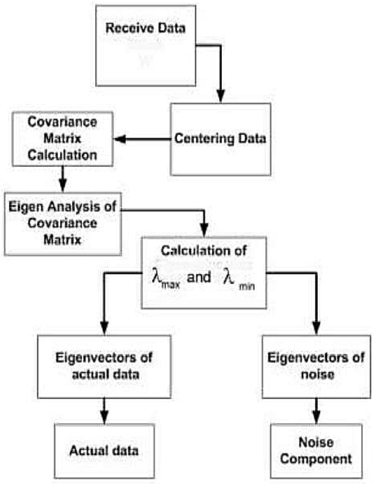
Receive Data
W
Centering Data
Covariance Matrix Calculation
Eigen Analysis of Covariance Matrix
Calculation of λ_{max} and λ_{min}
Eigenvectors of actual data
Eigenvectors of noise
Actual data
Noise Component

Figure 1. Flowchart of spectral filtering technique.

5 Random Matrix-Based Data Filtering

Suppose actual data S is perturbed by a noise random variable R to produce $W = S + R$. Let $w_i = s_i + r_i$, $i = 1, 2, \dots, m$, be m (perturbed) data points, each being a vector of n features. Thus the perturbed dataset, can be considered to be an $m \times n$ random matrix W, having n features and m instances. Our proposed filtering technique first calculates the covariance matrix of the perturbed data W. Using the distribution of eigenvalues of the covariance matrix, and the theory of random matrices, the covariance matrix of W is decomposed into a noise part and an actual data part. The eigenvectors corresponding to actual data are then used to reconstruct the actual data.

In the following section, we discuss some details of the filtering procedure. We first assume that the entire distribution $F_R(r)$ of the random noise R is known. Later, we discuss how the noise variance can be estimated from the eigenvalue distribution of the perturbed data.

5.1 Known Noise Variance

When the noise distribution $F_R(r)$ of R is completely known, the noise variance σ^2 is first calculated. Equation (2) is then used to calculate λ_{max} and λ_{min}. They provide the theoretical bounds of the eigenstates corresponding to noise. From the perturbed data, we compute the eigenvalues of its covariance matrix Y, say $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then we identify the noisy eigenstates $\lambda_i \leq \lambda_{i+1} \leq \cdots \leq \lambda_j$ such that $\lambda_i \geq \lambda_{min}$ and $\lambda_j \leq \lambda_{max}$. The remaining eigenstates are the eigenstates corresponding to actual data. Let, V_r = diag $(\lambda_i, \lambda_{i+1}, \ldots, \lambda_j)$ be the diagonal matrix with all noisy eigenvalues, and A_r be the matrix whose columns are eigenvectors corresponding to the eigenvalues in V_r. Similarly, let V_s be the eigenvalue matrix for the actual data part and A_s be the corresponding eigenvector matrix which is a $n \times k$ matrix ($k \leq n$). Based on these matrices, we decompose the covariance matrix Y into two parts, Y_s and Y_r with $Y = Y_s + Y_r$, where $Y_r = A_r V_r A_r'$, is the covariance matrix corresponding to random noise part, and $Y_s = A_s V_s A_s'$, is the covariance matrix corresponding to actual data part. An estimate $\hat{S}$ of the actual data S is obtained by projecting the data W on to the subspace spanned by the columns of A_s. In other words, $\hat{S} = W A_s A_s'$.

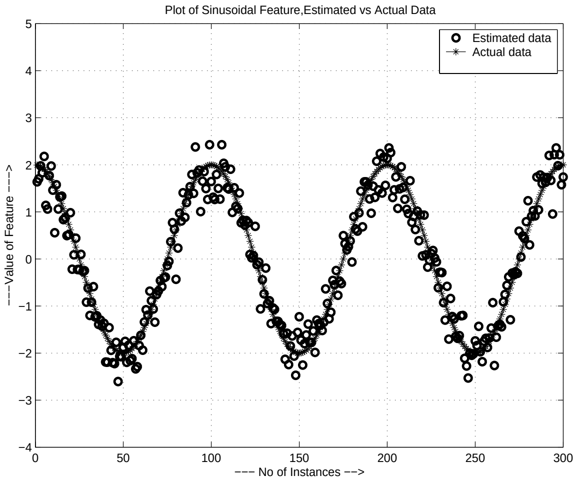

Figure 2. Estimation of original sinusoidal data with known random noise variance.

5.2 Unknown Noise Variance

When the noise variance σ^2 is unknown, we first estimate it using the perturbed data. The estimated noise variance is then used to filter the perturbed data. In order to estimate the noise variance σ^2 we first compute the eigenvalues of the covariance matrix Y of the perturbed data W. A histogram of the eigenvalue distribution is plotted and compared to that of the theoretical noise eigenvalue density function $f_Q(x)$ given in equation (2). Note that the density function $f_Q(x)$ depends on the variance σ^2. Typically, the theoretical density function $f_Q(x)$ is a good fit to the left portion of the histogram of the computed eigenvalues, corresponding to small eigenvalues. The larger eigenvalues that do not fit this theoretical density function correspond to the actual information part of the perturbed data. An iterative procedure is employed to obtain the value of σ that results in the best fit of $f_Q(x)$ to the observed histogram.

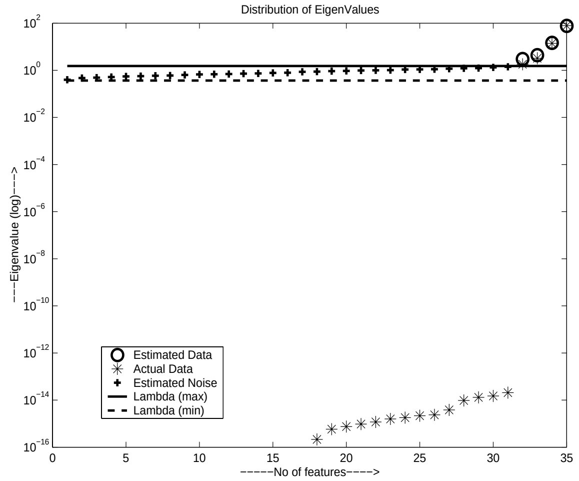

Figure 3. Distribution of eigenvalues of actual data , and estimated eigenvalues of random noise and actual data.

6 Experimental Results

Our proposed method is used on datasets of different sizes which have some trend in their values. The actual dataset is distorted by adding Gaussian noise (Normally distributed random numbers with zero mean and specific variance), and our proposed technique is applied to recover the actual data from the perturbed data with the knowledge of noise distribution (noise variance in particular). Experimental results show this method estimates the pattern and gives close estimation of individual values of actual data. Figure 2 shows one such estimation of data when the actual data has sinusoidal trend.

The distribution of eigenvalues shows (Figure 3) the method accurately distinguishes between noisy eigen values and eigenvalues corresponding to actual data. Note that the estimated eigenvalues of actual data is very close to eigenvalues of actual data and almost overlap with them above λ_{max}. The eigenvalues of actual data below the λ_{min} are of very small values and are negligible. Thus, even though there are no estimations corresponding to them, the estimation of actual data is fairly accurate.

We used a dataset of 300 instances and 20 features which has definite trend in its features. We added a Gaussian random variable with mean 0 and standard deviation $\sigma = 0.25$

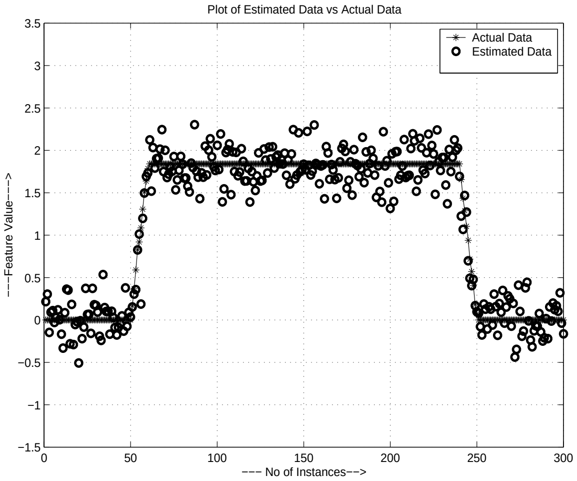

Figure 4. Estimated dataset preserves the 'Plateau' trend of original data.

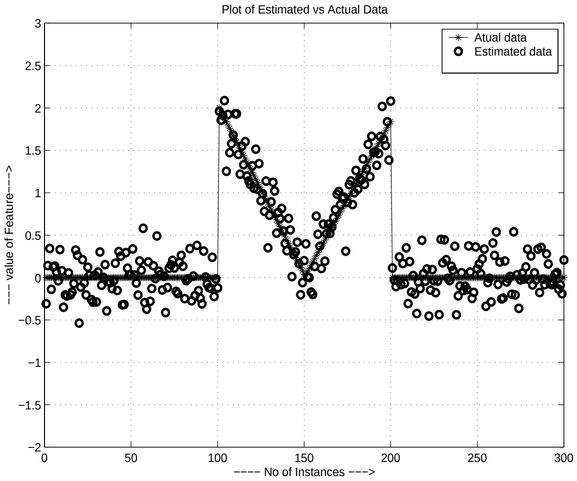

Figure 5. Estimated dataset preserves the 'Triangular' trend of original data.

to each data value of the actual dataset and applied our algorithm to recover the actual data from the distorted data with the known noise $\sigma = 0.25$. Figure 4, Figure 5 show estimation of dataset with different types of trends in their actual values. The actual dataset has trends like plateau and triangles. The estimated dataset preserves the trend and closely estimates individual values.

Quality of recovery depends upon relative noise content of the data. If the relative noise compared to actual dataset increases very much, the recovery method performs poorly. We define the term 'Signal-to-Noise Ratio' (SNR) to quantify the relative amount of noise added to actual data to perturb it.

$$\text{SNR} = \frac{\text{Value of Actual Data}}{\text{Value of Noise Added to the Data}}$$

As the noise added to the actual value increases, the SNR decreases. Our experiments show that this method predicts the actual data reasonably well up to a SNR value of 1.0 (i.e. 100% noise). The results shown in figures 2, 4, 5 are the case of mean SNR value nearly 2, i.e. 50% noise. As the SNR goes below 1, the estimation becomes too erroneous. Figure 6 shows the difference in estimation as the SNR increases from 1. The upper figure shows the estimation corresponding to 100% noise(mean SNR = 1), and the lower figure shows estimation corresponding to 16.67% noise (mean SNR = 6).

In case of unknown noise distribution, the method estimates the noise variance first. From the eigenvalues of covariance matrix of actual data, a histogram of the eigenvalue distribution is obtained, and this is compared with best possible theoretical density function given by Equation 2. The variance corresponding to the best fit gives the estimation of the noise variance.

To get the best estimation of variance, the algorithm estimates noise variance from the best fit curve several times. In each trial , the variance estimation algorithm starts with a very small variance value near zero, create the theoretically generated distribution and measures the mean square error between it and histogram of eigenvalues of actual data. It then increases variance by a small value, again computes the mean square error and compares it with the previous error to get the minimum error and corresponding variance. The algorithm does the said operation up-to a threshold value of variance, and stores of the variance corresponding to minimum mean square error between theoretically generated density function curve and histogram of eigenvalues of actual data.That value of variance is treated as the estimated value of noise variance for that particular trial.In our experiment, we used 100 such trials for each variance estimation. After the set of estimates are calculated from all trials, the distribution of estimated variances is checked for outliers in them. The mean μ_1 and standard deviation σ_1 of the estimates are calculated , and values lying outside the span $\mu_1 \pm 3\sigma_1$ are discarded. During each trial, if the algorithm does not get best fit within a predefined threshold value of variance, it stores that threshold value of variance as the estimation. These values are also treated as outliers at the end and are discarded.

After discarding the outlier estimations, an average of the rest of the estimates are taken to get the actual estimate of noise variance. We have noticed that discarding the outliers and taking average of the remaining number of estimate improves the estimation accuracy to a large extent. Figure 7 shows the theoretical density curve and distribu-

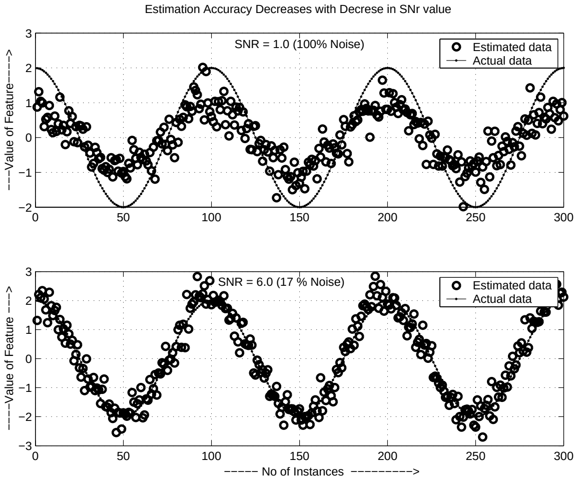

Figure 6. A higher noise content (low SNR)leads to less accurate estimation.SNR in upper figure is 1, while that for lower figure is 6.

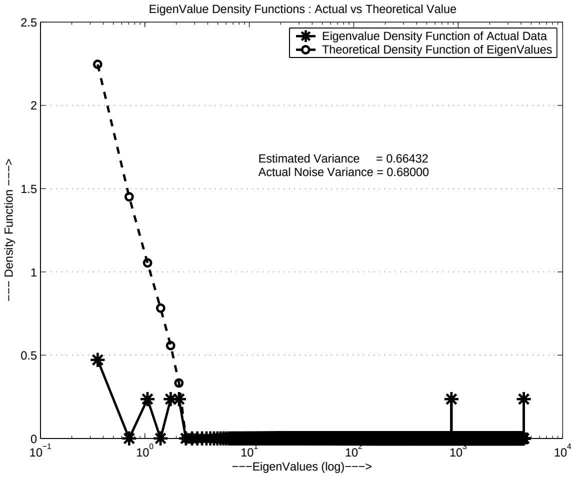

Figure 7. Theoretical density function and the actual distribution of eigenvalues.

tion of actual eigenvalues.The average over 100 estimates gives an estimated variance of 0.66432 where the actual noise variance is 0.68. Although not all the estimates are always so close, on average, the difference between the estimated and the true variances always remained within 10% of the actual variance in all our experiments.

Once the noise variance is estimated, the same technique is applied as before to estimate the original data. Figure 8 shows the estimation of actual data of a relatively small dataset with high SNR when distribution of noise is not known. Figure 9 displays the distribution of eigen values. The estimation of signal eigenvalues almost overlap with dominant eigenvalues of the actual dataset.

7 Conclusion and Future Work

Preserving privacy in data mining activities is a very important issue. This paper illustrates a noise filtering technique by which true data values can be estimated from the perturbed values (by random noise). This raises questions against the claim of preserving privacy by perturbing data with random numbers and disclosing the perturbed dataset as well as the probability distribution of the random number generator. The proposed approach works by comparing the empirically observed eigenvalue distribution of the given data with that of the known distribution of random matrices. The theoretically known values of upper and lower limits of the spectrum (eigenvalues) are used to identify the boundary between the eigen-states due to noise and that of the actual data.

This random matrix based approach to separating the information bearing and noisy eigen-states has potential computational advantages. Indeed, since the upper bound $\lambda_{\max}$ of the noisy eigenvalues is known a priori, one can easily use a suitable numerical technique (e.g., power method [10]) to compute just the few largest eigenvalues. Once these eigenvalues and corresponding eigenvectors are computed, one can obtain the actual-data-part of the covariance matrix, which can be subtracted off from the total covariance to isolate the noise-part of the covariance. The proposed approach is simple, and retrieves actual data with reasonable precision. For the datasets considered in this paper, our experimental results support this claim. So, the method of perturbing data with random number to hide their original value is not a very reliable method to preserve privacy.

This work leaves open the problem of coming up with methods which can actually preserve privacy without destroying statistical properties of the original dataset. Data mining application has the potential to reveal important trends in real-life data and use those trends to predict for the future. However a huge amount of sensitive dataset cannot be used just for the sake of preserving privacy. So a reliable data perturbation technique is necessary to use those sensitive datasets for data mining applications.

Acknowledgments

The authors acknowledge supports from the United States National Science Foundation CAREER award IIS-0093353, NASA (NRA) NAS2-37143, and TEDCO, Maryland Technology Development Center.

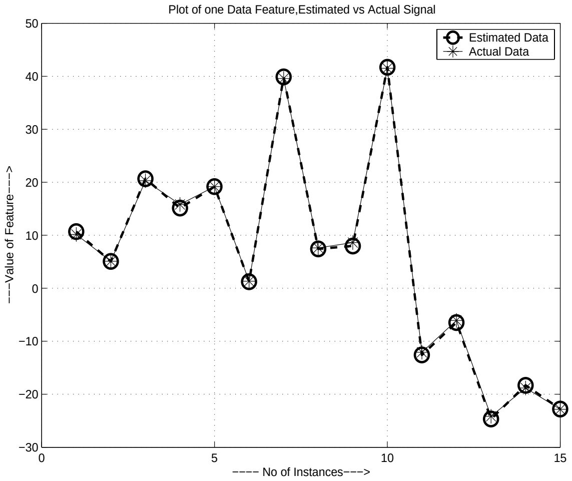

Figure 8. Estimation of actual data when the noise distribution is not known.

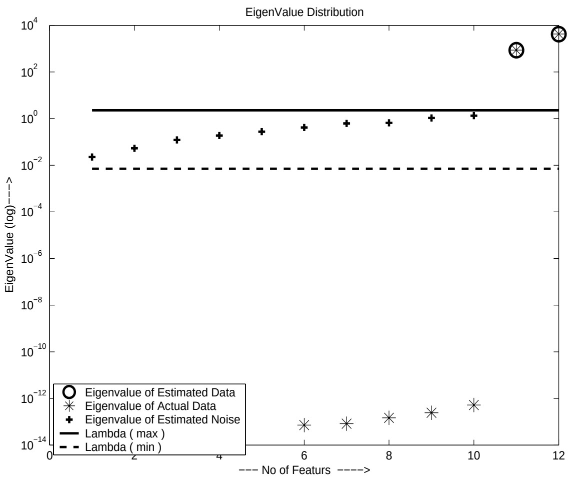

Figure 9. Distribution of eigenvalues for the estimations without the knowledge of noise distribution.

References

[1] Using unknowns to prevent discovery of association rules. 2001.

[2] Electricurrent technologies inc.netaccountability. 2002.

[3] E. Bertino A. Elmagarmid M. Ibrahim M. Atallah and V. Verykios. Disclosure limitation of sensitive rules. In *Proc. of IEEE Knowledge and Data Engineering Exchange Workshop (KDEX)*, 1999.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceeding of the ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, Texas, May 2000. ACM Press.

[5] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, August 1988.

[6] C. Clifton. Developing custom intrusion detection filters using data mining. In *2000 Military Communications International Symposium (MILCOM2000), Los Angeles, California*, 2000.

[7] Ahmed K. Elmagarmid Elena Dasseni, Vassilios S. Verykios and Elisa Bertino. Hiding association rules by using confidence and support. In *Lecture Notes in Computer Science*, page 369, 2001.

[8] S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, April 1980.

[9] U. Grenander and J. W. Silverstein. Spectral analysis of networks with random topologies. *SIAM Journal on Applied Mathematics*, 32(2):499–519, 1977.

[10] J. E. Jackson. *A User's Guide to Principal Components*. John Wiley, 1991.

[11] Tom Johnsten and Vijay V. Raghavan. Impact of decision-region based classification mining algorithms on database security. In *(IFIP) Workshop on Database Security*, pages 177–191, 1999.

[12] D. Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12:1–38, 1982.

[13] Murat Kantarcioglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *SIGMOD Workshop on DMKD*, Madison, WI, June 2002.

[14] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology CRYPTO 2000*, pages 36–54, August 2000.

[15] V. A. Marcenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR — Sbornik*, 1(4):457–483, 1967.

[16] M. L. Mehta. *Random Matrices*. Academic Press, London, 2 edition, 1991.

[17] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[18] J. W. Silverstein. On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 30(2):307–311, August 1989.

[19] J. W. Silverstein and P. L. Combettes. Signal detection via spectral theory of large dimensional random matrices. *IEEE Transactions on Signal Processing*, 40(8):2100–2105, 1992.

[20] J. F. Traub, Y. Yemini, and H. Woz'niakowski. The statistical security of a statistical database. *ACM Transactions on Database Systems (TODS)*, 9(4):672–679, 1984.

[21] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, CA, July 2002.

[22] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4):509–521, August 1988.

[23] Vassilios S. Verykios Yucel Saygin and Ahmed K. Elmagarmid. Privacy preserving association rule mining. In *RIDE*, 2002.