Military Technical College
Kobry El-Kobba,
Cairo, Egypt

11-th International Conference
on Aerospace Sciences &
Aviation Technology

A NOVEL PRIVACY PRESERVING DATA MINING ALGORITHM

Fahmy A. Aly* Fakhry M. Medhat** M. Ismail Hanafy*** El-Zeweidy M.Aly***

Abstract

In recent years, there have been privacy concerns over the increase of gathering personal data by various institutions and merchants over the Internet. There has been increasing interest in the problem of building accurate data mining models over aggregate data while protecting privacy at the level of individual records. One approach for this problem is to randomize the values in individual records, and only disclose the randomized values. This method is able to retain privacy while accessing the information implicit in the original attributes. The distribution of the original data set is important and estimating it is one of the goals of the data mining algorithms. In this paper, a novel privacy preserving data mining algorithm based on the use of Artificial Neural Network (ANN) is introduced. The ANN model is based on single layer neural network (adaptive linear neuron network (ADALINE)), and it is used to reconstruct the original distribution. The paper also introduces a comparative study with two of the most recent algorithms that handled this issue. Our empirical results show that the new algorithm can reconstruct the original data distribution with a very high degree of precision.

* Fahmy A. Aly	(Prof.) Dean of Faculty of Computers and information, Cairo University
** Fakhry M. Medhat	(Ass. Prof.) College of Engineering and Technology Arabic Academy for Science, Technology and Maritime Transport (AAST)
*** M. Ismail Hanafy.	(Ph.D.) Egyption Armed Forces
*** El-Zeweidy M.Aly	(Eng.) Egyption Armed Forces

1. INTRODUCTION

Data Mining is the process of efficient discovery of non-obvious valuable patterns (embedded facts and relationships) from a large collection of databases. Its goal is to create models for decision making that predicts future behavior based on analysis of past activities. The discovered data should not reveal secrets that are considered private for individuals or groups. The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms, which preserve user privacy. The conflict between privacy and data mining has lead to the devolvement of data mining algorithms that preserve the privacy of those whose personal data are collected and analyzed. The technical challenge is to provide security mechanisms for protecting the confidentiality of individual information used for knowledge discovery and data mining. More specifically, we need to develop techniques for replacing original data with data that approximately exhibits the same general patterns, but hide sensitive information; we need to develop mechanisms that will enable data owners to choose

an appropriate balance between privacy and precision in discovered patterns. Such techniques and mechanisms can lead to new privacy control systems to convert a given data set into a new one in such a way to preserve the general patterns from the original data set. The distribution of the original data set is important and estimating it is one of the goals of the data mining algorithms. The three algorithms that will be compared will be henceforth referred to as Algo-1, Algo-2 and Algo-3. Algo-1 is an iterative algorithm proposed in [7], Algo-2 is a Fourier-based estimate algorithm proposed in [1], and Algo-3 is a new proposed algorithm based on Artificial Neural Network (ANN). The information loss is utilized for each algorithm as a metric to study the performance of the privacy preserving data mining algorithms. While it is not possible to estimate the original data values in individual data records, the three algorithms will estimate the original distribution of the original data values. The distribution reconstruction process naturally leads to some loss of information which is acceptable in many practical situations. Previous work in privacy preserving data mining has addressed two broad approaches for privacy concerns namely, "Randomization approach" and "Cryptographic approach" in response to the conflict between privacy and data mining.
Section 2 discusses the conflict between privacy and data mining, the "Randomization" and the "Cryptographic" approaches.
Section 3, provides the current two data mining algorithms for distribution reconstruction.
Section 4, provides the new privacy preserving data mining algorithm which is based on Artificial Neural Network.
Section 5, provides analyze of the empirical results obtained by computer simulations for the three data mining algorithms.
Section 6, provides conclusions and discussions.

2. PRIVACY CONCERNS, CONFLICT BETWEEN PRIVACY AND DATA MINING

The main privacy issue is that secrets that are considered private for individuals or groups should not be revealed. The concept of privacy is neither clearly understood, nor easily defined [10]. An advanced concept of privacy suggested by Moor in [6], called the "control/restricted access theory". In Moor's definition, the notion of a situation is emphasized to distinguish the loss of privacy from a violation of privacy.
According to [6], in the process of determining whether certain information is private or public, it is neither the kind of information, nor the content of information itself, rather, it is the situation or context in which the information is used that can determine it. Taking Moor's example, we can see that at private colleges' faculty salary schedules are often kept confidential, whereas at larger state colleges faculty salaries are sometimes open in public. Along this line, Moor's control/restricted access theory can be applied to the process of determining private vs. public concerns of privacy protection.
The balance between privacy and the need to explore large volumes of data for pattern discovery is a matter of concern. There are different views of the Knowledge Discovery and Data Mining (KDDM) experts, and different issues related to the conflict between privacy and data mining. KDDM discover patterns that classify individuals into categories.
Approaches for privacy in KDDM have only recently been considered, however, none have been applied seriously for KDDM. All the privacy protection methods proposed

for KDDM are well known and applied in the context of statistical databases. There, methods have been developed to guard against the disclosure of individual data while satisfying requests for aggregate statistical information [12].

2.1 Approaches to resolve the conflict between privacy and Data Mining

In this section, two approaches to resolve the conflict between privacy and Data Mining will be discussed. The first approach is the randomization approach; the second one is the cryptographic approach, then a comparison between the two approaches and the scenarios of use of each of them is addressed.

- **Randomization Approach**

The idea of the "Randomization approach" is that you can take data from a population, add a random variable to it and then recover important characteristics from this perturbed data. This method to preserve the privacy of data is called "Value distortion" [9].

The "Randomized approach" relies on the notion that one's personal data can be protected by being scrambled or randomized prior to being communicated, "Randomizing people's information as they enter it can result in data nearly as good as the real thing, if it's subjected to some post-processing". The level of that randomization, and the resulting privacy, depends on the software settings [9].

For instance, instead of recording the answer "41" to a curious question like "How old are you?", the software automatically adds a random number of years within a specified range, say minus 30 to plus 30, to the answer. No record of initial answers is kept. For example, Susan enters her age as 30. It's randomized to 42. Mary enters her age as 34, which is randomized to 28. This continues for every person who enters his/her age. The resulting aggregate randomized data is processed and "corrected" by the software. Then, using a series of mathematical guesses based partly on how the initial data was randomized, the program gradually reconstructs a realistic distribution of the age groups that responded, how many people were 20 to 25, say, or 40 to 45. Demographic information like this might be of great interest to a company in quest of 25-year-olds to buy its sports cars or computer games [9].

By "adding random values to true values, the S/W can reconstruct a distribution that is very close to the actual one. After collecting all the randomized data for a large number of users, the data mining software would use the randomized distribution to reconstruct what the true distribution might have been. When you do this for 10,000 answers, the overall distribution is likely to be accurate [9].This technique relies on two facts:

- Users are not equally protective of all values in their records. Thus, users may be willing to provide modified values of certain fields by the use of a (publicly known) perturbing random distribution. This modified value may be generated using a custom code or a browser plug-in.
- Data mining problems do not necessarily require individual records, but only distributions. Since the perturbing distribution is known, it can be used to reconstruct aggregate distributions, i.e. the probability distribution of the data set. In many cases, data mining algorithms can be developed which use the probability distributions rather than

individual records. An example of a classification algorithm which uses such aggregate information is discussed in [7].

- **Cryptographic Approach**

The second privacy approach is the cryptographic approach. In this approach the problem is addressed from a cryptographic standpoint in [4] and [13-14] where data mining computations among several parties are performed on the combined data sets of the parties without revealing each party's data to the other parties. A key problem that arises in any huge collection of data is that the level of confidentiality. The need for secrecy is sometimes due to law, or can be motivated by business interests. However, sometimes there can be mutual gain by sharing of data. A key utility of large databases today is research, whether it is scientific or economic and market oriented. Consider a scenario in which two or more parties owning confidential databases wish to run a data-mining algorithm on the union of their databases without revealing any unnecessary information. For example, consider separate medical institutions that wish to conduct a joint research while preserving the privacy of their patients. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party.

The three algorithms that will be addressed in this paper are based on the randomization approach. The distributed computing scenario is outside the scope of this paper.

3. PRIVACY PRESERVING DATA MINING ALGORITHMS

In this sections the current two data mining algorithms for distribution reconstruction that address the issue of data mining while privacy of data is kept, will be discussed, their implementation results will be verified and commented in the next section.
The two algorithms are based on the randomization technique, which addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining. In this technique adding a random value from a known distribution perturbs the individual data in a large data set, and the data mining algorithm does not know the true value of the data. In these applications, the distribution of the original data set is important and estimating it is one of the goals of the data mining algorithm. This method is able to retain privacy while accessing the information implicit in the original attributes.

3.1 Problem Definition

For the concept of using the randomization approach to protect privacy to be useful, we need to be able to reconstruct the original data distribution from the randomized data (note that, we reconstruct distributions, not original values).
As the problem that will be discussed is the same for the three algorithms, so it will be defined once in this section. The basic problem can be abstracted into the following mathematical problem.
Consider a set of n original data values x_1, x_2, x_n. These are modeled in [7] as n independent values, each drawn from the same data distribution of the random variable X. In order to create the perturbation (to hide these data values), we

generate n independent values y_1 , y_2 , ... y_n, each with the same distribution as the random variable Y . We assume that X and Y are independent. Thus, the perturbed values of the data are given by $z_1 = x_1 + y_1, \ldots z_n = x_n + y_n$.
In order to protect privacy, only the perturbed values are provided rather than the original data. Given these perturbed values z_i and the density function $f_Y(y)$ of Y, the goal is to estimate the density function $f_X(x)$ of X. In the example of the internet survey, x_i corresponds to the participants' answers, y_i correspond to the perturbations generated and z_i correspond to the perturbed answers which are sent to the server for collection.

3.2 Implementation of the first algorithm (Algo-1)

Given the perturbed values z_i and the (publicly known) density function $f_Y(y)$ for Y, the goal is to estimate the density function $f_X(x)$ of X. Algo-1 is an iterative algorithm that was proposed in [7] to estimate the density function $f_X(x)$ for X.

- **Reconstruction Algorithm**

1. Start with a uniform distribution as an estimate for the original data distribution at the initial state.
2. Start iteration at j =0 over the available randomized data set of length n.
3. Compute the updated density function of X according to equation number (1).
4. Increase the counter by one step increment and repeat step 3 until the density function reach a steady state (until a stopping criterion is met). Stop when the difference between successive estimates of the original distribution becomes very small (1% of the threshold of mean square test).
   - f_X^0 := Uniform distribution.
   - j := 0 // Iteration number.
   - Repeat

$$f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y((x_i+y_i)-a)f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i+y_i)-a)f_X^j(a)} \qquad (1)$$

   - J:= J + 1
   - Until (stopping criterion met).

3.3 Implementation of the second algorithm (Algo-2)

The Fourier-based estimate algorithm (Algo-2) that has been implemented was based on an algorithm proposed by [1]. The algorithm is based on a Fourier series method to compute in one step a good initial estimate of the distribution in order to reduce the number of iterations or eliminate the iterative step completely that was used in other algorithms such as Algo-1, and so to reduce the amount of computation in the estimation algorithm.
The detailed equations that show how the Fourier coefficients of f_X can be computed in one step to generate an initial estimate of f_X are described in appendix-1.

♦ **Reconstruction Algorithm**

It is assumed that the distribution of the randomizer f_Y is known and there are n records of perturbed data Z.
The steps of the algorithm are as follows:

1. Compute the coefficient matrix of the randomizer Y as in equation (10) using the given f_Y
2. Compute the Fourier series coefficient a_i, b_i according to equation (9).
3. Compute the estimate of the probability density function (pdf) f_X of the real data according to equation (2).

4. NEW PRIVACY PRESERVING DATA MINING ALGORITHM

The third algorithm is a novel privacy preserving data mining algorithm based on the use of ANN. The new algorithm is also based on the randomization approach, which addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining. An ANN model based on single layer neural network (adaptive linear neuron network (ADALINE)) is used to reconstruct the original distribution. Fig.1 shows the architecture of the linear network.
The transfer function of the ADALINE is linear which permits its output to take on any value. The learning rule used here is the Least Mean Square (LMS) learning rule, which minimizes the mean square error.
The adaptive linear system responds to changes in its environment as it is operating. Linear networks that are adjusted at each time step based on new input and target vectors can find weights and biases that minimize the network's sum-squared error for recent input and target vectors. Such networks are often used for error cancellation problems [8].

4.1 Modes of operation

1- Training Mode

During this mode, the adaptive linear network is trained on examples of correct behavior. For the purpose of training the network, the input to the network will be the perturbed samples Z_i , the well known perturbing density function f_Y, and the correct behavior density function f_X. The output of the network is the estimated density function $\hat{f}_X$ which is compared with f_X. The LMS algorithm adjusts the weights and biases of the ADALINE so as to minimize the mean square error (m.s.e.). The mean square error (m.s.e.) is defined as:

$$m.s.e = \frac{1}{N}\sum_{i=1}^{N}(\hat{f}_X(i) - f_X(i))^2$$

Where N is the number of bins
When the mean square error (m.s.e.) is minimized, then the optimum network is obtained.

2- Test Mode

During the test mode, we apply the perturbed samples Z_i and the perturbing density function f_Y, as the input to the network. The output of the pre-optimized network $\hat{f}_X$ will be a good estimate for the unknown original distribution f_X.

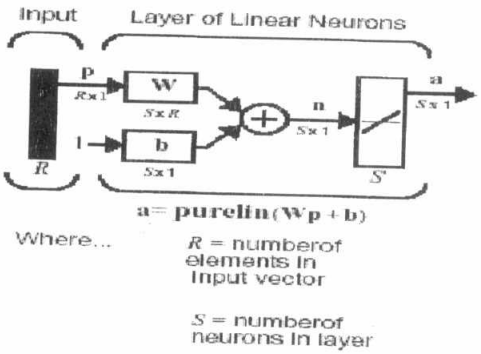

Fig. 1 linear network architecture

4.2 Information loss metric

By adding perturbation, we lose precision in estimating f_X, the density function of X. In [3], this is referred to as information loss and a metric is proposed to quantify this. The metric, was denoted as I, and defined as the expected value of the statistical difference (also called variation distance [15] or Kolmogorov distance [16]) between the original distribution of X and the estimated distribution. Given the perturbed values z_1 , z_2 , …,z_n, it is (in general) not possible to reconstruct the original density function $f_X(x)$ with an arbitrary precision. The greater the variance of the perturbation, the lower the precision in estimating $f_X(x)$. We refer to the lack of precision in estimating $f_X(x)$ as information loss. This metric is zero for perfect reconstruction. Note that this metric depends on k, the number of bins used in estimating the distribution.

To measure the information loss, we propose the metric of the mean square error (MSE) which is calculated as the square of the difference between the estimated distribution and the original distribution.

$$MSE = E[(\hat{f}_X - f_X)^2]$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{f}_X(i) - f_X(i))^2$$

Where N .....is the number of bins,

$\hat{f}_X$ ... is the estimated density function,

f_X ... is the original density function.

5 EMPIRICAL RESULTS

In this section, the reconstruction accuracy of the three algorithms is presented. In many cases, data mining algorithms can be developed to use the probability distributions rather than individual records. The distribution reconstruction process naturally leads to some loss of information, which is acceptable in many practical situations.

Information loss is an important metric to capture the amount of data in an individual record leaked to the data mining algorithm and the reliability of the estimate respectively. One of the design goals of such privacy preserving data mining algorithms is to derive algorithms, which can have a small information loss. It is clear that privacy loss is small when the perturbation is large and vice versa.

5.1 Methodology

We compare the reconstruction accuracy of the three algorithms with respect to the information loss metric (by calculating the mean square error between the estimated distribution and the original one).

To illustrate the estimation method and measure the performance of the three algorithms, a MATLAB program is implemented for each of them with the same test bed as follows:

1- The original data is a Bimodal Distribution which is generated using two Gaussian random variables (using a pseudo random number generator) with means m1, m2 (0, 0.17) and variances σ_1^2 and σ_2^2 (1, 1). Then, at random, equally probable samples from them are selected. Although, most of the real phenomena's can be interpreted as Gaussian random variable (population salaries, students degrees...), we didn't use it as it is well known that the Gaussian random variable achieve the maximum uncertainty compared to other random variables. We used the bimodal distribution through the paper as indicated in [1]. Fig.2 shows the original "Bimodal" data distribution that will be used during our experimentation.

2- Perturbed training data is generated using
   a. Uniform Distribution in the range [0, 1].
   b. Gaussian distribution with mean (0) and variance (1).

3- All results involving randomization were averaged over 10 runs.

5.2 Case study

Fig.2 shows the original "Bimodal" data distribution using a sample of 100,000 data points and k = 128 bins. Fig.3 shows the noise distribution "Uniform in the range [0, 1]", that will be used to perturb the original distribution indicated in Fig.2. In Fig.4 we have shown the case of the perturbed data distribution, using the original unperturbed "Bimodal distribution", and the perturbed distribution with Uniform noise. Fig.5 presents the original and the perturbed data distribution, using Uniform noise. In Fig.6, we have shown the original distribution, the perturbed distribution, the estimated distribution, and the estimation error by the help of the Algo-3 reconstruction algorithm using Uniform noise.

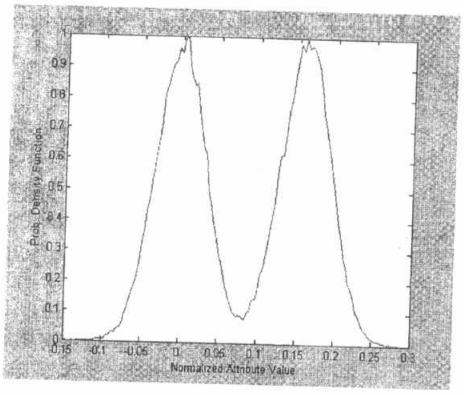

Fig.2 Original Bimodal Data Distribution

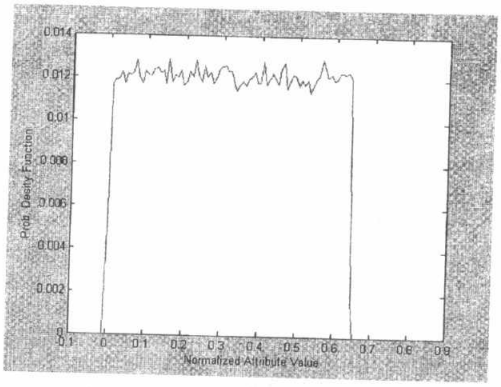

Fig.3 Perturbation Data Distribution "Uniform"

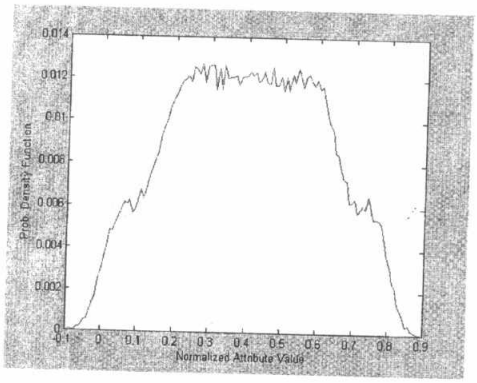

Fig.4 Perturbed Data Distribution using "Uniform" Noise.

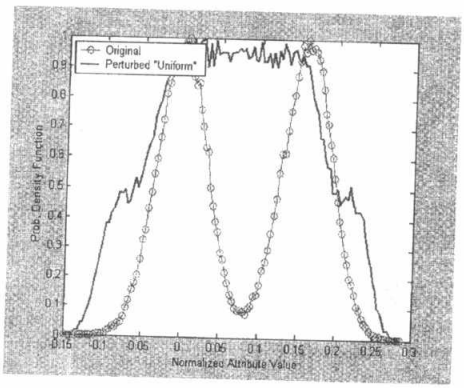

Fig.5 Original Bimodal distribution, perturbed distribution using Uniform noise

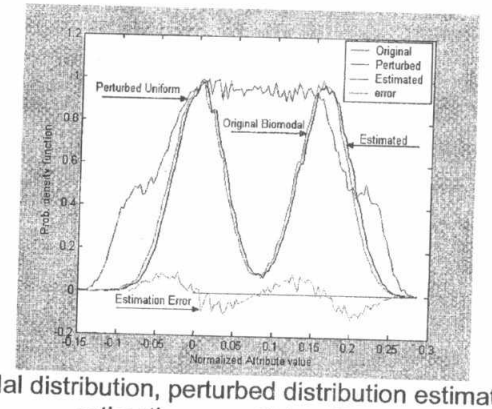

Fig.6 Original Bimodal distribution, perturbed distribution estimated distribution, and estimation error Using Algo-3.

One can see that the proposed algorithm Algo-3 was able to reconstruct the original distribution with some acceptable tolerance. The MSE between the original distribution and the reconstructed one is further studied in the next section.

5.3 Comparison of the proposed algorithm and the traditional ones

In this section, we present some interesting trends of the privacy-preserving reconstruction algorithms. It turns out that the three algorithms are almost competitive in their performance. However, the (Algo-2, Algo-3) reconstruction algorithms are able to reconstruct the data distribution more effectively than the (Algo-1) algorithm.
In Figures 7, 8 and 9, we have illustrated one such case in which we reconstructed a Bimodal distribution with the use of the Algo-1, the Algo-2 and the Algo-3 algorithms. In this case, the original data contains 50,000 points (Bimodal distribution with means m1, m2 (0, 0.17) and variances σ_1^2 and σ_2^2 (1, 1)) respectively. We added a uniformly distributed noise in the range [0, 1]) in order to perturb the original data. The unperturbed distribution is computed using k = 50 bins. Using the Algo-2 algorithm, the density function is estimated using Fourier coefficients a_i, b_i = 1, 2,....12 and forced to nonnegative.
In Fig.7, we have shown the reconstructed distribution obtained by the Algo-1 algorithm. The corresponding level of information loss is 11,37%.
On the other hand, in Fig.8 we have shown the reconstructed distribution obtained with the use of the Algo-2 algorithm. The information loss level is 7,6%. We see that the Fourier-based estimate algorithm "Algo-2" which is computed in a much shorter time than the Algo-1 is better in quality than the Algo-1 algorithm.

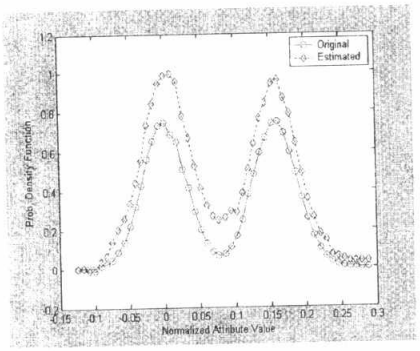

Fig.7 Original-Estimated Distributions using Algo-1 with Uniform Noise

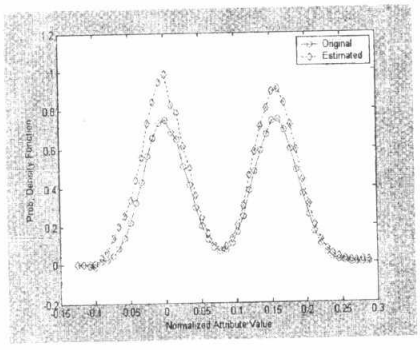

Fig.8 Original-Estimated Distributions using Algo-2 with Uniform Noise

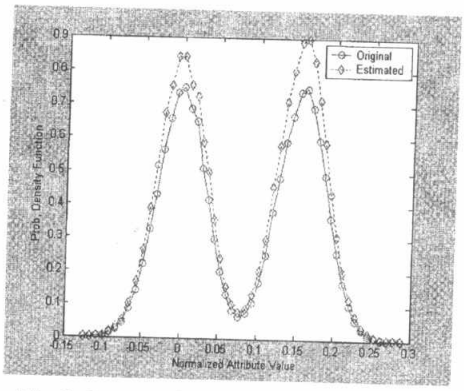

Fig.9 Original-Estimated Distributions using Algo-3 with Uniform Noise

In [3] it was also observed via simulation that when f_X is Gaussian, the choice of whether f_Y is uniform or Gaussian has little effect on the information loss, whereas when f_X is uniform, the information loss for uniform f_Y is smaller than the information loss for Gaussian f_Y. This observation can be explained by noting that the spectrum of a Gaussian distribution drops off faster than the spectrum of a uniform distribution. This means that it is easier to reconstruct f_X when f_Y is uniform and it is harder to reconstruct f_X when f_Y is Gaussian, especially when f_X is uniform.
The information loss level for the Algo-3 algorithm, is only 3,7%. The corresponding distribution is illustrated in Fig.9.
In Figures 10, 11, and 12, we have shown another case in which we reconstructed a Bimodal distribution with the help of the Algo-1, the Algo-2 and the Algo-3 algorithms respectively. The perturbing distribution is Gaussian with mean (0) and variance σ^2 (0.0134). The original bimodal distribution is generated using a sample of 50,000 data points. In this case, we found that the information loss level for Algo-1 was 22,01%, while for Algo-2 was 13,17% and for Algo-3 was as low as 6,7%,
This difference is shown in Fig.10 as the amount of mismatch with the original distribution as compared to the mismatch produced by the Algo-2 (shown in Fig.11) and also with the mismatch produced by the Algo-3 algorithm (shown in Fig.12).

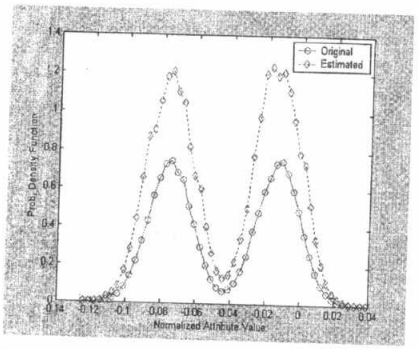

Fig.10 Original-Estimated Distributions using Algo-1 with Gaussian Noise

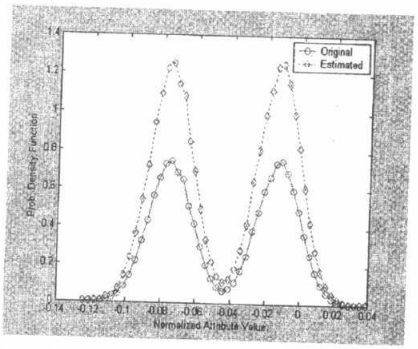

Fig.11 Original-estimated Distributions using Algo-2 with Gaussian Noise

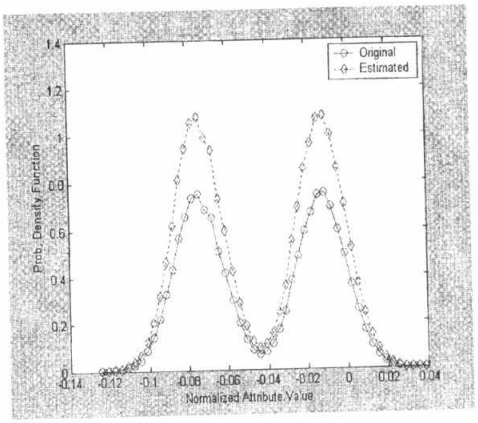

Fig.12 Original-Estimated Distributions using Algo-3 with Gaussian

In Fig.13, we plot the information loss with increasing the variance of the perturbing distribution using the Algo-3 algorithm. The variance of the perturbed data is changed between 0.1 to 1 of the average value of the perturbed attribute.
These results are presented for three different combinations of the original and perturbing distributions, using the Bimodal distribution as the original one, Uniform and Gaussian distributions as the perturbing distributions. The range of the uniform noise was [0, 1] whereas the variance of the Gaussian noise was (0.0134). In each case there were 50,000 data points. As expected, the amount of information loss grows with the level of perturbation. The greater amount of loss of information with increased perturbation (as shown in Fig.13) comes at the advantage of high privacy level. Moreover, one can see that the information loss is higher in case of Gaussian perturbation than that one for uniform perturbation. This is because; the Gaussian perturbation has higher degree of uncertainty than the uniform one.

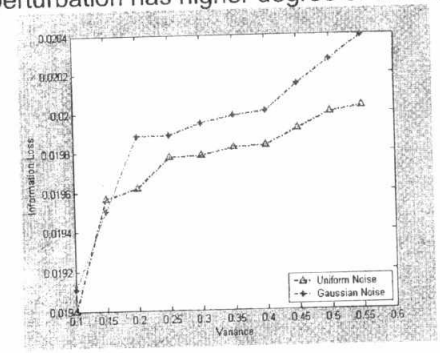

Fig.13 Information Loss---Variance using Algo-3

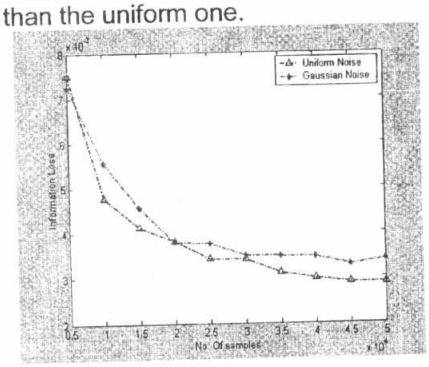

Fig.14 Information Loss --- No. of Samples using Algo-3

In Fig.14, we have shown the behavior of the Algo-3 reconstruction algorithm with increasing the number of points in the data. This curve corresponds to the case when

the original distribution is Bimodal and the perturbing distributions were Uniform and Gaussian.

It is noticed that when there are a large number of data points, the amount of information loss decreases.

Our empirical results show the following trends of privacy-preserving data mining algorithms:

(1) With increasing perturbation, the privacy level increases, but the effectiveness of reconstruction algorithms decreases.
(2) With increasing the amount of data available, the Algo-3 reconstruction algorithm is able to approximate the original distribution to a very high degree of precision.

6. Conclusion and discussion

In this paper, three privacy-preserving data mining algorithms were discussed. One of them is proposed as a novel privacy preserving data mining scheme based on the use of ANN, where the density function of the original data set can be estimated using adaptive linear neuron network (ADALINE). The results showed that it is competitive in its performance with the more complicated iterative procedures that have been proposed in the past (Algo-1). This algorithm provides a robust estimate of the original distribution.

We qualified the relative effectiveness of different perturbing distributions using information-loss metric. Our tests also demonstrate that when the data is large then Algo-3 can reconstruct the data distribution with a very small information loss. Furthermore, because the estimation algorithm "Algo-2" is essentially a summation of the vectors Z_i's, it computes in one-step a good initial estimate of the distribution in order to reduce the number of iterations in Algo-1 or eliminate the iterative step completely. This is in contrast to the iterative algorithm (Algo-1), where all the vectors Z_i's are needed at each iteration. From the results, the proposed algorithm provides the minimum information loss compared to the other two algorithms.

REFERENCES

[1] Chai Wah Wu, "Privacy Preserving Data Mining: a Signal Processing Perspective and a Simple Data Perturbation Protocol" IBM Research Report RC22815 (W0306-040) June 9, (2003)

[2] B.W. Silverman, Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, Chapman and Hall, (1986).

[3] D. Agrawal and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Symposiun on Principles of Database Systems,(2001).

[4] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, (2002).

[5] N.N. Cencov, "Evaluation of an unknown distribution density from observations" Soviet Math., vol.3,pp.1559-1562, (1962).

[6] Moor, James H. "Toward a Theory of Privacy in the Information Age," Computers and Society, vol.27, 3, pp.27-32, (1997).

[7] R. Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," in Proc. of the ACM SIGMOD Conference on Management of Data, pp. 439{450, ACM Press, May 2000.

[8] Widrow B. and S. D. Sterns, *Adaptive Signal Processing*, New York: Prentice-Hall 1985.

[9] Rakesh Agrawal , IBM Scientists Rely on the Principle of Uncertainty To Develop Web-Privacy", May 30, (2002). http://www.krcollin@us.ibm.com,

[10] Tavani, Herman T. (2000) Privacy and the Internet. Paper presented at the Ethics & Technology Conference, June 5, 1999. [Online] Available at: http://www.bc.edu/bc_org/avp/law/st_org/iptf/commentary/

[11] S.C. Schwartz, "Estimation of probability density by an orthogonal series," Annals of Mathematical Statistics, vol.38,pp. 1261-1265, (1967).

[12] Vladimir Estivill-Castro, Ljiljana Brankovic and David L. Dowe Privacy in Data Mining, August (1999).

[13] W. Du and M. J. Atallah, "Privacy-preserving cooperative scientific computation," in 14th IEEE Computer Security Foundations Workshop, (2001).

[14] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", Crypto 2000, August (2000).

[15] A. Sahai and S. Vadhan, "Manipulating statistical difference," in Randomization Methods in Algorithm Design (DIMACS Workshop, December 1997) (P. Pardalos, S. Rajasekaran, and J. Rolim, eds.), vol. 43 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 251{270, American Mathematical Society, 1999.

[16] M. Nielsen and I. Chuang, Quantum computation and quantum information. Cambridge University Press, 2000.

APPENDIX-1

In this appendix, we show how the Fourier coefficients of f_X can be computed in one step to generate an initial estimate of f_X. Estimating the Fourier coefficients of f_X belongs to the class of orthogonal series estimators [5], [11]. This estimate can be further refined using the iterative methods of [7], [3]. When the one-step estimate is close to f_X, the number of iterations needed in the refinement step is smaller than in [7], [3]. Assume that the data is properly scaled so that the support of X is a subset of [0, 1]. We want to express f_X as a Fourier series defined on the interval [0,1]:

$$f_X(x) = a_0 + \sum_{i=1}^{\infty} = a_i \sin(2\pi i x) + \sum_{i=1}^{\infty} b_i \cos(2\pi i x) \quad (2)$$

The Fourier series coefficients of the function $f_X(x)$ are given by

$$a_i = 2\int \sin(2\pi i x) f_X(x) dx \quad (3)$$

$$b_i = 2\int \cos(2\pi i x) f_X(x) dx \quad (4)$$

Since f_X is a probability density function, $a_0 = 1$. As $\sqrt{2}\sin(2\pi X)$ and $\sqrt{2}\cos(2\pi X)$ are Ortho-normal in the interval [0,1], it follows that

$$a_i = 2E[\sin(2\pi i X)] \quad (5)$$

$$b_i = 2E[\cos(2\pi i X)] \quad (6)$$

Where E denotes the mathematical expectation. It is easy to verify that

$$E[\sin(2\pi i Z)] = E[\sin(2\pi i(X+Y))]$$
$$= E[\sin(2\pi i X)]E[\cos(2\pi i Y)] + E[\cos(2\pi i X)]E[\sin(2\pi i Y)] \quad (7)$$

$$E[\cos(2\pi i Z)] = E[\cos(2\pi i(X+Y))]$$
$$= E[\cos(2\pi i X)]E[\cos(2\pi i Y)] - E[\sin(2\pi i X)]E[\sin(2\pi i Y)] \quad (8)$$

Therefore

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} = 2\begin{pmatrix} E[\cos(2\pi i X)] \\ E[\sin(2\pi i X)] \end{pmatrix} = 2A_i^{-1}\begin{pmatrix} E[\sin(2\pi i Z)] \\ E[\cos(2\pi i Z)] \end{pmatrix} \quad (9)$$

Where

$$A_i = \begin{pmatrix} E[\cos(2\pi i Y)] & E[\sin(2\pi i Y)] \\ -E[\sin(2\pi i Y)] & E[\cos(2\pi i Y)] \end{pmatrix} \quad (10)$$

Since f_Y is known, A_i^{-1} can be computed in advance and $E[\sin(2\pi i Z)]$, $E[\cos(2\pi i Z)]$ are estimated as $\frac{1}{n}\sum_{j=1}^{n}\sin(2\pi i z_j)$ and $\frac{1}{n}\sum_{j=1}^{n}\cos(2\pi i z_j)$ respectively.

One of the drawbacks of this method is that it works as long as Ai is not close to be singular for the coefficients that we are interested in. In other words, this method works well, i.e., has a smaller information loss, if f_Y has higher frequency components than f_X. Note that the estimated density function is independent of the number of bins used in the iterative algorithms of [7], [3]. As is common in orthogonal series estimators, the Fourier coefficients need to be smoothed e.g., via finite truncation of the Fourier series or by weighting the coefficients [2].

INSTRUMENTATION & ELECTRICAL SYSTEMS

CONTENTS