A Proposed Model to Allow Data Mining Classification Avoiding Privacy Concerns

EL-Zeweidy M. Aly (Ph.D.)

Higher institute of computer and information technology (Shourouk Academy)
Melzeweidy1@yahoo.com

Abstract

Data Mining aims to discover hidden facts that exist in the databases and data warehouses. The discovered data should not reveal secrets that are considered private for individuals or groups. 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.

This paper introduces the privacy concerns and the obvious conflict between privacy and data mining. Then, two approaches to resolve this conflict are introduced, namely: the randomization approach and the cryptographic approach.

We consider the case of performing data mining classification for randomized data. Two proposed algorithms for data mining classification of randomized data ,with high accuracy compared to classification algorithms for non perturbed data, based on Bayes rules will be introduced (Step-Class, and Global-Decision).

These two algorithms are experimentally tested to measure the classification accuracy of each of them. Our empirical results show that the Step-Class algorithm has better performance results (classification accuracy ratio) than the Global decision algorithm.

Keywords

Knowledge Discovery and Data Mining (KDDM), Bayes classifiers, privacy

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 [2-3]. The distribution of the original data set is important and estimating it is one of the goals of the data mining algorithms. Two new Bayesian classifier algorithms (Step-Class, and Global-Decision) that can be used to classify perturbed data with high accuracy compared to classification algorithms for non perturbed data are presented. These two algorithms are experimentally tested to measure the classification accuracy of each of them. The results showed that the Step-Class algorithm has better performance results (classification accuracy ratio) than the Global-Decision algorithm

Previous work for estimating the original distribution can be found in [1,2, 14]. 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 [2]. The paper is organized to include five sections. Section 2 discusses the conflict between privacy and data mining, the “Randomization” and the “Cryptographic” approaches. Section 3, provides the design and implementation of randomized data classification algorithms, then the analysis of the Bayesian classification model for the two proposed algorithms is introduced. Section 4, provides analyze of the empirical results obtained by computer simulations for the two proposed algorithms. Section 5, 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. An advanced concept of privacy suggested by Moor in [6,10], called the “control/restricted access theory”.

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 [11].

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]. 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 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. More details about distributed computing scenarios can be found in [4], [12,13].

The two 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. Design and implementation of randomized data classification algorithms

This section provides the two proposed Bayesian classifier algorithms. We consider the case of performing data mining classification for randomized data. The two proposed algorithms for data mining classification of randomized data are based on Bayes rules and will be henceforth referred to as step-class algorithm and Global-decision algorithm. Figure (1) shows the post-processing stage which includes the implementation of the two classification algorithms. The analytical solution of the Bayesian classification model is given in the next section.

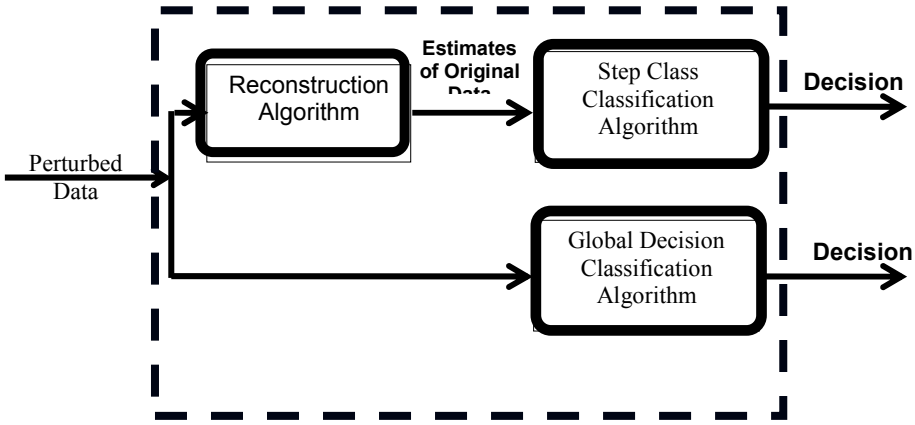

Figure (1) Classification algorithms

3.1 Bayesian classification Model

In this section we will introduce the analytical solution of the Bayesian classification model of the two classification algorithms that will be implemented and verified. The purpose of Bayesian classification is to classify a randomized set of data as correct as if the real data are available to the classifier. The main elements of the Bayesian classifier are:

- A set of attributes A_i ; $i = 1,2,.....N$ which describe a population samples. Each attribute could take a set of values M_i ; $i = 1,2,.....N$ that the real values of the attributes are

$A_{11}, A_{12}, \ldots . A_{1M_1}$, $A_{21}, A_{22}, \ldots . A_{2M_2}$, ... $A_{N1}, A_{N2}, \ldots . A_{NM_N}$

These data are represented in the form of records. Each record contains a single value from each attribute.

- A set of predefined classes $C_1, C_2, C_3, \ldots . C_M$. Each class contains those records whose attributes satisfy predetermined conditions.
- It is assumed that an arbitrary record describes a single point in the sample space. Thus, the dimension of a record is N-dimensional vector $R_l = (A_{1l_1} A_{2l_2} A_{3l_3} \ldots . A_{Nl_N})$ *where* $1 \le l_i \le M_i$ $; i = 1,2, \ldots . . N$

The problem of data mining while privacy is preserved; is that; the available records do not contain the real values of the attributes. Rather a randomized version of these attributes values are presented at the input of the data mining algorithm. To make the problem clear, assume that the record R_l of the sample space takes the first values of each attribute $R_l = (A_{11} A_{21} \ldots . . . . . A_{N1})$. Due to privacy concerns; this record will not appear at the input of the Bayesian classifier; instead a randomized version Z will appear, such that $Z = (Z_1 Z_2 \ldots . . . . . Z_N)$ where $Z_i = A_{i1} + y_i$ $; i = 1,2,3 \ldots . . N$ where y_i is a random variable added to the original value of the attribute A_i for privacy concerns of data. It is assumed that the probability density function (pdf) $f_y(\alpha)$ is known to the classifier. The main task of the Bayesian classifier is to attach the record Z to a class C_i correctly as well as if the real record R_l is available to the classifier.

It is worth to mention that the set of random variables $y_i ; i = 1,2, \ldots , N$ are statistically independent. This does not mean that we consider only the naïve Bayesian network, but we consider also the Bayesian belief network, since the classes are defined on a joint description of all the considered attributes.

To simplify the problem we consider in the first part of this section that each record contains only one attribute. Moreover, this attribute A_1 takes only two values A_{11}, A_{12} with probabilities p , 1 – p. Clearly, we could define only two classes C_1, C_2 for such type of data. Section 2 considers a little bit complicated problem, that each record contains a single attribute A_1 , but this attribute could take M_l values $A_{11}, A_{12}, \ldots . A_{1M_1}$. Clearly, we can define a set of M classes $C_1, C_2, C_3, \ldots . . C_M$ *where* $M \le M_l$. Finally, the analysis of the most complicated case of N attributes with M classes will be analyzed in section 2.

3.1.1 Single attribute with two classes

We consider the case of one attribute, A_1 which has two values $\{ A_{11}, A_{12} \}$, the probability of occurrence of them is p, q=1 – p respectively. The corresponding classes could be simply defined as C_1 *or* C_2 . The problem of data mining is concerned with classification of a randomized version of this attribute and attaching it to one of pre-known classes C_1 *or* C_2 that $A_1 = A_{11} \Rightarrow C = C_1 ;$ $A_1 = A_{12} \Rightarrow C = C_2$.

Consider that the original data is $A_{1i}; i = 1,2$ and the randomization will be done through addition of a random variable Y, so the observed randomized data will be Z that is given by:

$$Z = A_{1i} + Y; \quad i = 1,2$$

The random variable Y is considered as Gaussian random variable with probability density function $f_Y(y)$, zero mean (m) and variance (σ_y^2). Addition of this random variable to the original attribute is used for hiding the real value of the attribute. The probability density function of Y is given by

$$f_Y(y) = \frac{1}{\sigma_y \sqrt{2\pi}} e^{\frac{-y^2}{2\sigma_y^2}} \qquad \text{...(1)}$$

It is clear that the added random variable Y is Continuous, meanwhile the original attribute A is discrete, and then Z is also continuous.
We can get the distribution of the observation Z based on the actual value of the attribute as follows:

$$F_{Z|A_{1i}}(z) = pr[Z \le z | A_1 = A_{1i}] = \Pr[Y + A_{1i} < z] = pr[Y \le z - A_{1i}] \quad ; i = 1,2 \qquad \text{...(2)}$$

By taking the derivative of both sides with respect to z we get

$$f_{Z|A_{1i}}(z|A = A_{1i}) = f_y(z - A_{1i}) \qquad \text{...(3)}$$

$$F_{Z|A_{11}}(z|A_1 = A_{11}) = F_y(z - A_{11})$$

$$F_{Z|A_{12}}(z|A_1 = A_{12}) = F_y(z - A_{12})$$

Since Y is a Gaussian Random variable, we can write that

$$f_{Z|A_{11}}(z|A = A_{11}) = \frac{1}{\sigma_y \sqrt{2\pi}} e^{\frac{-(z-A_{11})^2}{2\sigma_y^2}} \qquad \text{...(4)}$$

$$f_{Z|A_{12}}(z|A = A_{12}) = \frac{1}{\sigma_y \sqrt{2\pi}} e^{\frac{-(z-A_{12})^2}{2\sigma_y^2}} \qquad \text{...(5)}$$

Note that equation (4, 5) is Gaussian distribution with mean $A_{1i}, i = 1,2$ and variance. σ_y^2

It is desired to decide whether the observation Z belongs to class. C_1 *or* C_2 The maximum a posterior probability rule [8], which minimizes the probability of decision error, performs the following operations.

The decision is $\hat{C} = C_1$ *iff*
$pr[A_{11}|Z = z] > pr[A_{12}|Z = z]$

These posteriori probabilities $pr(A_{1i}|Z=z)$ is not known to the classifier, however we can use Baye's theorem, that

$$P(X|Y)=\frac{P(Y|X)P(X)}{P(Y)} \qquad \text{...(6)}$$

Consequently the decision rule becomes:

$\hat{C}= C_1$ *iff* $\frac{pr[Z=z|A_{11}]P(A_{11})}{P(Z=z)} > \frac{pr[Z=z|A_{12}]P(A_{12})}{P(Z=z)}$

Thus

$\hat{C}= C_1$ *iff* $pF_{Z|A_{11}}(z|A_{11}) > qF_{Z|A_{12}}(z|A_{12})$

else

$$\hat{C}= C_2 \qquad \text{...(7)}$$

Since, the distribution function of a random variable is an increasing function, so we can replace it by its derivatives.

By differentiating both sides of equation (7) with respect to z we get

$$\hat{C}= C_1 \text{ iff } \quad pf_{Z|A_{11}}(Z|A_{11}) > qf_{Z|A_{12}}(z|A_{12}) \qquad \text{...(8)}$$

Substitute in (8) from 4,5 we get

$\hat{C}= C_1$ *iff* $\frac{p}{\sigma_y\sqrt{2\pi}}e^{\frac{-(z-A_{11})^2}{2\sigma_y^2}} > \frac{q}{\sigma_y\sqrt{2\pi}}e^{\frac{-(z-A_{12})^2}{2\sigma_y^2}}$

$\hat{C}= C_1$ *iff* $\exp[\frac{-1}{2\sigma_y^2}]\left[(z-A_{11})^2-(z-A_{12})^2\right] > \frac{q}{p}$

The exponential function is monotonic increasing function, so we can take the natural log for both sides that results in

$\hat{C}= C_1$ *iff* $\frac{-1}{2\sigma_y^2}\left[(z-A_{11})^2-(z-A_{12})^2\right] > \ln\frac{q}{p}$

$$Let \quad \alpha \triangleq \ln\left(\frac{q}{p}\right) \qquad \text{...(9)}$$

$\therefore \hat{C}= C_1$ *iff* $(z-A_{11})^2-(z-A_{12})^2 < 2\sigma_y^2\alpha$

By simple mathematical manipulation, we get the final form of the decision rule that will be as follows:

$\hat{C}= C_1$ *iff* $Z < T_h$

and

$\hat{C} = C_2 \qquad iff \qquad Z < T_h$

Where the threshold value T_h is related to the attribute values A_{11}, A_{12} and the disturbance variance by the relation

$$T_h = \frac{\left(A_{11}^2 - A_{12}^2\right) - \alpha}{2(A_{11} - A_{12})} \qquad \textbf{...(10)}$$

Clearly, the threshold is a constant value and can be computed in advance. To simplify the problem, let $p = q = \frac{1}{2}$ then $\alpha = 0$ and the threshold value will be $T_h = \frac{A_{11} + A_{12}}{2}$

Actually the problem of data mining and classification is treated here as a two hypothesis test. The observation space of Z is divided by the threshold T_h into two disjoint regions. Clearly the function of the classifier is easy to decide the category of any record by observing its location in the classified regions. It must be noted that during the implementation of data hiding, the choice of the standard deviation σ_y (strength of randomization) determines the level of the privacy. As σ_y increase, the privacy increases. To illustrate this, assume that an attribute " A " takes only two values (A_{11} , A_{12}) with equal probability. Then the average value of this attribute is. $\frac{A_{11} + A_{12}}{2}$ We chose σ to achieve a relative strength 'R' between the attribute and the corresponding randomizer such that the attribute to randomize ratio $R = \frac{E[A_1^2]}{E[Y^2]} = \frac{E[A_1^2]}{\sigma^2}$

During the simulation, we chose $\sigma = r(\frac{A_{11} + A_{12}}{2})$ where $0 < r < 1$

One can see graphically as shown in Figure (2) that as R decreases, σ increases leading to the increase of the dispersion around the value of the attribute ($A_{11} or A_{12}$) which increases the privacy. Consequently, the data mining algorithm will suffer from randomization.

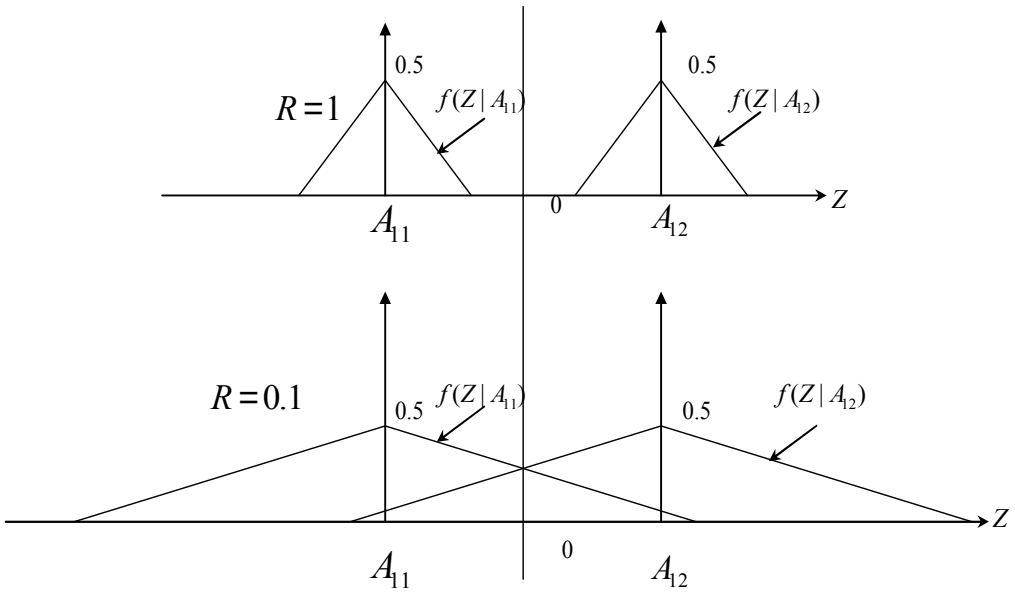

Figure (2) Effect of attribute to randomize ratio "R" on the available data

3.1.2 Single attribute with multi values and multi classes

Here we assume that each record consists of a single attribute utilized for data mining. This attribute has M distinct values. The weight or the occurrence of each value is related to the total population by a probability distribution. Thus, assume that the attribute A takes the values $A_1, A_2, A_3, \ldots A_M$ with corresponding probability $p_1, p_2, p_3, \ldots p_M$ respectively. The probability density function of such attribute is given by the Dirac distribution, that

$$f_A(a) = \sum_{i=1}^{n} p_i \delta(a - A_i) \qquad \text{Where} \qquad \delta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases}$$

It is assumed that each value A_i of the considered attribute is corresponding to a distinct class C_i. Usually to achieve privacy, a random variable y with known probability density function, $f_Y(y)$ is added to the real value of the attribute A. Thus, a classifier or a data mining algorithm will observe a randomized version value Z for the attribute, which is given by

$$Z = A_i + Y \qquad ; i = 1,2,3,.....M$$

It must be noted that the data mining algorithm or the classifier knows, in advance the following:

I. The set of all possible real values of the attribute.
$A_1, A_2, A_3, \ldots A_M$

II. The probability density function of the randomizer Y .

The task of the data mining algorithm is to associate the observed value $Z = Z_i$ to a possible class C_l, $1 \le l \le M$ based on the above knowledge.

Usually to increase the privacy (uncertainty), the random variable Y is chosen to be Gaussian with zero mean and variance σ_y^2. It is well known that, the Gaussian random variable has the highest degree of uncertainty; consequently it provides a high degree of privacy.

Since the attribute A_1, has M_1 set of possible values $A_{11}, A_{12}, A_{13}, \ldots A_{1M_1}$, and the randomizer Y has continuous distribution, the observation Z_l has a continuous value. This situation could be represented graphically as shown in figure (3). The infinite set of points along the straight line Z represents the continuous random observation Z, while the finite set of dots represent the values of the considered attribute A.

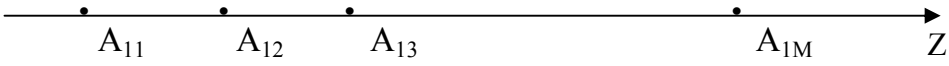

Figure (3) The observation space Z with the values of the attribute A_1

The classifier; using the pre-mentioned data in (I, II) divides the observation space Z into M distinct regions. $R_1, R_2, R_3, \ldots R_M$ as shown in figure (4). If the observation Z_l is within the region R_K, the classifier will decide that $C = C_k$.

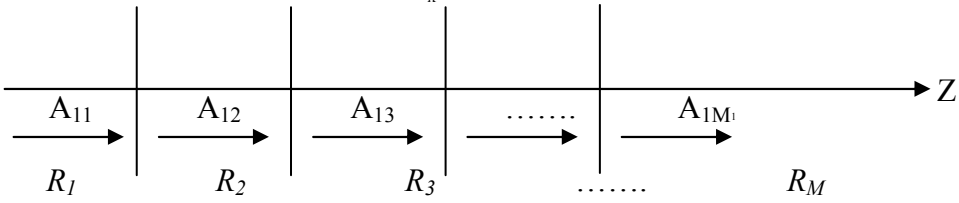

Figure (4) The observation space Z is divided into M distinct regions

The boundaries of the decision regions are determined by the Bayes decision rule given below:

$$Z \in R_k \quad iff \quad p_k f_{Z|A_{1k}}(z|A_{1k}) > p_i f_{Z|A_{1l}}(z|A_{1l})$$

; $l = 1, 2, \ldots M$

3.1.3 Global-Decision Based on Baye's Rule

Usually data mining algorithms are built on the basis of multiple attributes. In this section, we will introduce the global decision based on Baye's rule.

Assume that the randomized vector $\underline{Y}$ is used to hide the real data in the record $\underline{X}$. Clearly $\underline{Y}$ has M components, each of them represents a random

variable y_i with certain probability density function, and we denote it by. $f_{y_i}(\alpha)$ Furthermore, we assume that the randomized variable is added to the original data which gives us a sufficient degree of privacy; meanwhile it facilitates the possibility to reconstruct the distribution of the original data. The randomized record is given by: $\underline{Z} = \underline{X} + \underline{Y}$ where $\underline{X} == (A_{1x}, A_{2x}, A_{3x}, ...A_{Nx})$; $\underline{Y} == (y_1, y_2, y_3, ...y_N)$; $\underline{Z} == (z_1, z_2, z_3, ...z_N)$.

It must be noted that, randomization of each attribute value X_i of the record X is carried out by adding a random variable Y_i, independent from all other attributes. Mathematically, it is expressed as follows:

$$f_{y_i y_j}(\alpha, \beta) = f_{y_i}(\alpha) f_{y_j}(\beta) \quad \forall \quad y_i \neq y_j$$

This approach of data randomization has the advantage of controlling the randomization process of all different attributes. Clearly some attributes have a special property, and must be highly randomized to hide this special property. On the other hand, other attributes do not require this high degree of randomization. There are two reasons for that; first they are common parameters among all the population space, so that discovering their real values by non-authorized persons will not violate the system privacy. Second, these types of attributes represent a high priority key in the classification and processing the data records, consequently high randomization of this type of attributes will lead directly to wrong decision or incorrect classification.

However, in our problem, the data mining is concerned with decision making on the randomized record $\underline{Z}$. It could be noted that this situation represents repetitions of the previous case. In many literatures the problem of N attributes is solved in an iterative method, that each attribute in the record is treated independent of the others. The distribution function of each attribute is reconstructed at first. The values of this attribute are discretized and the resultant records are treated as the real records to make the classification process. We denote such approach as the step-class method.

In global decision method we will not make decision for individual attributes, rather we will make single decision for the overall randomized record in one step. Let us first derive the joint probability density function of the randomization vector $\underline{Y}$.

The probability distribution function of the random vector $\underline{Y}$ is given by, [8]

$$F_{\underline{Y}}(\underline{y}) = pr\left[Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_N \leq y_N\right] \qquad \textbf{...(11)}$$

Since it is assumed that; Y_i is independent from all other attributes; this lead to

$$F_{\underline{Y}}(\underline{y}) = F_{y_1}(y_1), F_{y_2}(y_2), ..., F_{y_N}(y_N)$$

Assume that each attribute is randomized by Gaussian random variable Y_i that is $N(0,\sigma_i^2)$ then

$$f_{Y_i}(y_i)=\frac{1}{\sigma_i\sqrt{2\pi}}e^{\frac{-y_i^2}{2\sigma_i^2}} \qquad \text{...(12)}$$

The joint probability density function of the random vector $\underline{Y}$ is given by

$$F_{\underline{Y}}(\underline{y})=\frac{1}{\left(\prod_{i=1}^{M}\sigma_i\right)(2\pi)^{M/2}}\exp\frac{-1}{2}\sum_{i=1}^{M}\left(\frac{y_i}{\sigma_i}\right)^2 \qquad \text{...(13)}$$

For simplicity we assume that we have a finite set of classes, that the real record $\underline{X}$ belongs to it, $S=\{C_1,C_2,C_3,.....C_M\}$. It is required to associate the randomized record $\underline{Z}$ to one class of S say C_k such that; if the real record $\underline{X}$ is the available one to the data mining algorithm, it will be associated to the same class C_k.

An optimum decision rule on Z will select $\hat{C}$ such that $P(\hat{C}\neq C)$ is minimum.
Without loss of generality, assume for the moment, that the real class of the record $\underline{X}$ is the class C_k then the observed record will be

$$\underline{Z}=\underline{X}^{(k)}+\underline{Y} \qquad \text{...(14)}$$

Of course the optimum decision rule is given by:

$$\hat{C}=C_k \quad \textit{iff} \quad pr(C_k|\underline{Z})>pr(C_l|\underline{Z}) \quad ;1\le l\le M,\quad l\neq k \qquad \ldots(15)$$

Apply Baye's rule as in single attribute case then

Decide that $\hat{C}=C_k$ *iff*

$$pr\left(\underline{Z}/\underline{X}^{(k)}\right)P\left(\underline{X}^{k}\right)>pr\left(\underline{Z}/\underline{X}^{(l)}\right)P\left(\underline{X}^{l}\right) \qquad \text{...(16)}$$

For simplicity we assume that all set of classes are equally probable, this implies

$$P(X^{(1)})=P(X^{(2)})=.....(X^{(N)})=1/N \qquad \text{...(17)}$$

The decision rule will be

$\hat{C}=C_k$ *iff*

$$pr\left[\underline{Z}/X^{(k)}\right]>pr\left[\underline{Z}/X^{(l)}\right] \qquad \text{...(18)}$$

Now we find $pr\left[\underline{Z}/X^{(k)}\right]$ in terms of the known pdf $F_y^{(n)}$

$$pr\left[\underline{Z} \le \underline{Z}_0 / X^{(k)}\right] = pr\left[\underline{Y} \le \underline{Z}_0 - X^{(k)}\right]$$
$$pr\left[\underline{Z} \le \underline{Z}_0 \ / X^{(k)}\right] = F_{\underline{Y}}\left(\underline{Z}_0 - X^{(k)}\right) \quad \text{...(19)}$$

$$\widehat{C} = C_k \quad iff$$
$$F_{\underline{Y}}\left(\underline{Z}_0 - \underline{X}^{(k)}\right) \rangle \ F_{\underline{Y}}\left(\underline{Z}_0 - \underline{X}^{(l)}\right) \quad \text{...(20)}$$

Replacing the distribution function with the joint density of $\underline{Y}$ we get

$$f_{\underline{Y}}\left(\underline{Z} - \underline{X}^{(k)}\right) \rangle \ f_{\underline{Y}}\left(\underline{Z} - \underline{X}^{(l)}\right) \quad \text{...(21)}$$

Substituting from (12) we get

$$\exp\left[\frac{-1}{2}\sum_{i=1}^{M}\frac{\left(\underline{Z}_0(i) - \underline{X}_k(i)\right)^2}{2\sigma_i^2}\right] \quad \rangle \quad \exp\left[\frac{-1}{2}\sum_{i=1}^{M}\frac{\left(\underline{Z}_0(i) - \underline{X}_l(i)\right)^2}{2\sigma_i^2}\right] \quad \text{...(22)}$$

Taking the natural log for both sides we get:

$$\left[\frac{-1}{2}\sum_{i=1}^{M}\frac{\left(\underline{Z}_0(i) - \underline{X}_k(i)\right)^2}{2\sigma_i^2}\right] \quad \rangle \quad \left[\frac{-1}{2}\sum_{i=1}^{M}\frac{\left(\underline{Z}_0(i) - \underline{X}_l(i)\right)^2}{2\sigma_i^2}\right] \quad \text{...(23)}$$

Let us denote the Euclidian distance between vector X_k and the observed vector $\underline{Z}_0$ as d_k where

$$d_k = \sqrt{\sum_{i=1}^{M}\left(Z_0(i) - X_k(i)\right)}^{\,2}$$

Then $\widehat{C} = C_k$ *iff* d_k is the minimum distance for k=1,2,…. M

4 Performance evaluation of the two proposed algorithms through computer simulation

In this section, the empirical results of the two proposed algorithms (Step-Class and Global-Decision) for classification of randomized data as an implementation of Bayes theory will be presented, evaluated and commented. At the beginning, the general methodology that will be applicable for the two algorithms will be defined, and then the results of each algorithm will be presented and discussed.

4.1 Methodology

The implementation of the two proposed algorithms was done to compare them at the same platform and at the same environment. To help in comparing both algorithms the following assumptions were made:

1. Both algorithms were implemented on the same hardware (same computer with the same processor and memory).
2. Both algorithms were implemented on the same operating system (Windows XP professional).
3. No additional processes were running in the background.
4. No scheduled programs were running.
5. No Screen saver was chosen.

We compare the classification accuracy (ratio of correct decision) of the step-class and the Global-decision algorithms (by calculating the mean square error between the decision based on the real record $\underline{X}$ and that one based on the randomized record $\underline{Z}$. The error function is defined as

$$\varepsilon(\underline{X}^{(k)}) = 1 \quad if \quad \hat{C}(\underline{Z}) \neq \hat{C}(\underline{X}^{(k)}) \qquad \text{... (24)}$$

$$\varepsilon(\underline{X}^{(k)}) = 0 \quad if \quad \hat{C}(\underline{Z}) = \hat{C}(\underline{X}^{(k)})$$

Where $\hat{C}(\underline{Z})$ is the class of the record $\underline{X}^{(k)}$ decided based on the randomized observation $\underline{Z}$ and $\hat{C}(\underline{X}^{(k)})$ is the class of the record $\underline{X}^{(k)}$ decided based on the real record $\underline{X}^{(k)}$. The ratio of the correct decision is computed based on extremely large number of records for both algorithms.

Clearly, we want to come close in accuracy to the original classification as possible.

During the simulation, a sample of 100,000 records is used. Perturbed data is generated using both Uniform and Gaussian distribution.

To recall how privacy could be achieved during the simulation, a random variable Y with known probability density function $f_Y(y)$ is added to the real value of the attribute A. Thus, the classifier will observe a randomized version value Z of the attribute, which is given by $Z = X + Y$.

The performance of the data mining algorithm is evaluated by a quantitive measure called the ratio of correct decision as function of the statistical average of the set of values that each attribute takes relative to the strength of the randomizer. The ratio of the strength of the attribute relative to the added randomized variable is defined as:

$$R = \frac{E[A_i^2]}{E[Y_i^2]} \qquad \text{...(25)}$$

Where $E[A_i^2]$ is the mean square value of the attribute A_i, and $E[Y_i^2]$ is the mean square value (variance) of the randomized variable Y_i added to the attribute A_i.

4.2 Pre-processing of input data sets

During the simulation, only categorical attributes are considered (i.e. each attribute has a finite set of values). Each record is constructed from five predictor attributes and one class attribute. The predictor attributes are described in table (1) and their distributions are shown in Figure (5). The classification functions that

have been used are described in table (2). The selected set of attributes and functions are just to verify our proposed algorithms. The analytical solution as mentioned before is suitable to deal with N attributes A_i ; $i = 1,2,.....N$ which describe a population sample. Each attribute could take a set of values M_i ; $i = 1,2,.....N$ that the real values of the attributes are

$$A_{11}, A_{12}, ... A_{1M_1}, \quad A_{21}, A_{22}, ... A_{2M_2}, ... \quad A_{N1}, A_{N2}, ... A_{NM_N}$$

These data are represented in form of records. Each record contains a single value from each attribute. A set of predefined classes $C_1, C_2, C_3, ... C_M$ is also defined. Each class contains those records whose attributes satisfy predetermined conditions.

Attribute #	Description
A1	Could take a value of (0 or 1).
A2	Could take a value of (0 , 1 or 2).
A3	Could take a value of (0 , 1, 2 or 3).
A4	Could take a value of (0 , 1, 2, 3 or 4).
A5	Could take a value of (0 , 1, 2, 3, 4 or 5).

Table (1) Attributes description

Function	Description	Class
Function 1	(A3 <= 2 & A4<2 & A5 <=3)	C1
Function 2	(A3<=2 & A4<2 & A5>3)	C2
Function 3	(A3 <= 2 & A4>= 2 & A5<=3)	C3
Function 4	(A3<= 2 & A4>= 2 & A5>3)	C4
Function 5	(A3> 2 & A4<2 & A5<=3)	C5
Function 6	(A3<=2 & A4<2 & A5>3)	C6
Function 7	(A3>2 & A4>= 2 & A5<=3)	C7
Function 8	(A3>2 & A4>= 2 & A5 >3)	C8

Table (2) Functions description

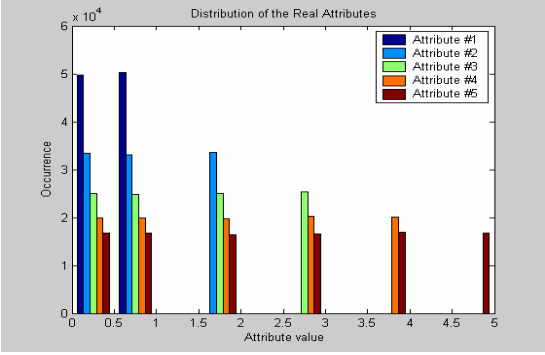

Figure (5) Distribution of the original attributes

4.3 Evaluation of the step-class algorithm

In this section we present a case study for the first algorithm (Step-class) which is modeled as shown in Figure (6). The model is based on the Naïve Bayesian Classifier that makes the assumption of class conditional independence that is, given a class label of a sample; the values of the attributes are conditionally independent of one another.

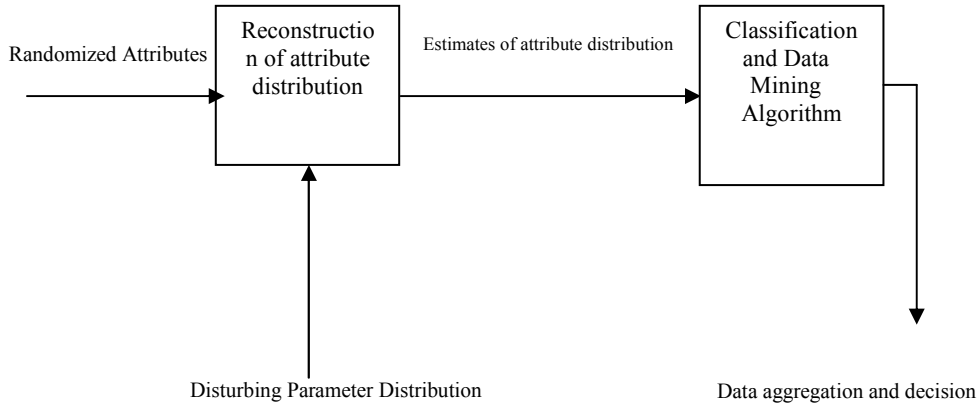

Figure (6) Step-class model block diagram

In this model the individual attributes are hidden by adding random values to the true values with a predetermined ratio. The software would access only the randomized values and the parameters of randomization. Based only on this information; the software could reconstruct a close approximation of the true distribution for each attribute. This continues for all attributes. This process is shown in Figure (7). After reconstructing the original distribution of each attribute, the classification is performed for each corrected record based on Bayesian classifier to predict the class label of each record for which the class label is missing or unknown according to the predefined set of classes.

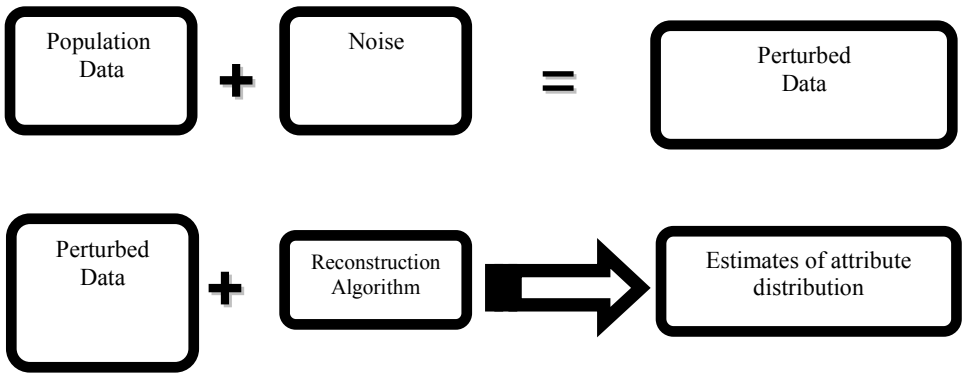

Figure (7) Reconstruction of attribute distribution

To show the effect of changing the strength of the randomizer relative to the attribute value, we assume different values for (R), the ratio of the attribute strength relative to the randomizer as shown in figures. We vary the variance of the randomizer by he same value for each attribute and evaluate the ratio of the correct decisions. The experiment is performed using Gaussian and Uniform randomizers to perturb the original data. The results are presented in Figure (8) to Figure (13). Figures (8) to (10) show the original, randomized, and estimated data distribution for different values of "R" using "Gaussian" randomizer for a sample of 100,000 data records. Figures (11) to (13) show the original, randomized, and estimated data distribution for different values of "R" using "Uniform" randomizer for a sample of 100,000 data records. It is clear from these figures that as the variance of the randomizer goes high, the distribution of the attribute is highly disturbed, and many clusters appear which do not express the real distribution of this attribute. Consequently, the reconstruction algorithm fails to determine the true distribution as shown in Figure (8), Figure (11) where R = -10 db (i.e. the strength of the randomizer is 10 times the strength of the attribute value). For middle values of the randomizer (R=10db where, the strength of the randomizer is one tenth the strength of the attribute value), the reconstruction algorithm could determine a good estimate for the original data distribution as shown in Figure (9), Figure(12). As the variance of the randomizer goes low, the reconstruction algorithm succeeded to determine a better estimate of the true distribution as shown in Figure (10), Figure(13) where R = 20 db (i.e. the strength of the attribute value is 100 times the strength of the randomizer).

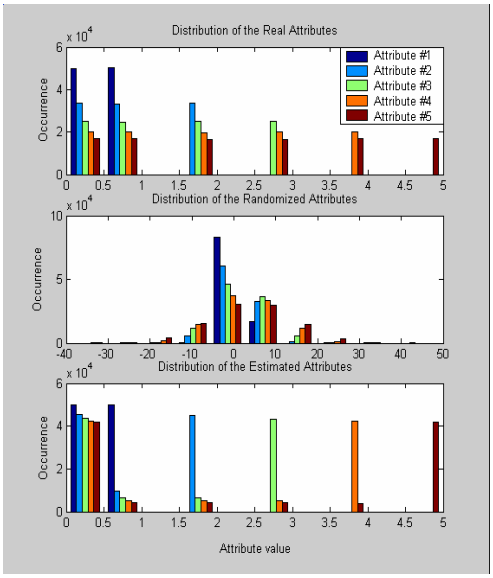

Figure (8) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = -10dB using “Gaussian” Noise

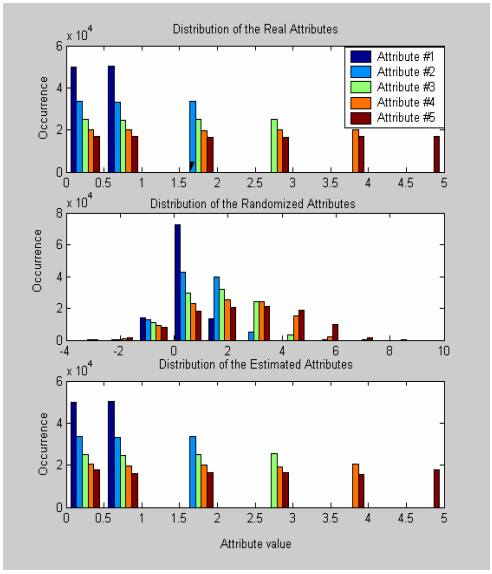

Figure (9) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = 10dB using “Gaussian” Noise

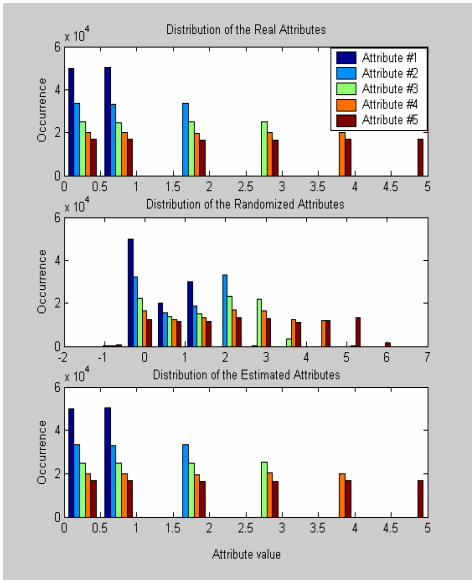

Figure (10) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = 20dB using “Gaussian” Noise

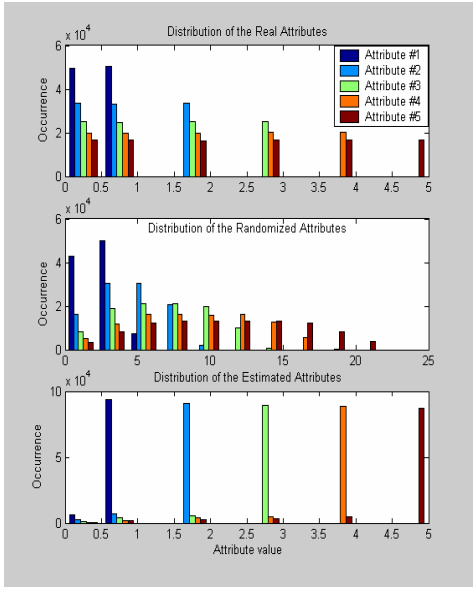

Figure (11) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = -10dB using “Uniform” Noise

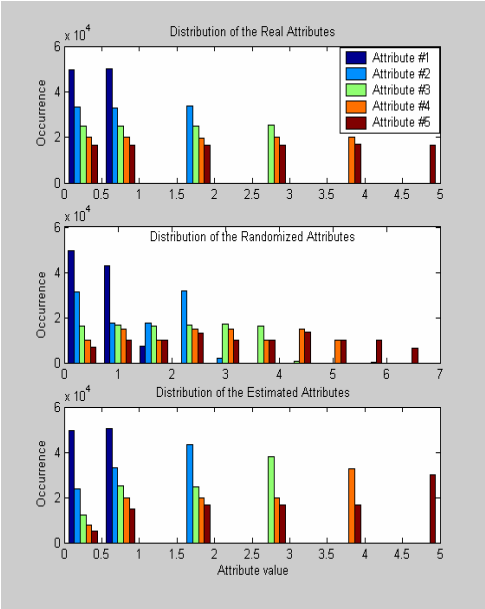

Figure (12) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = 10dB using "Uniform" Noise

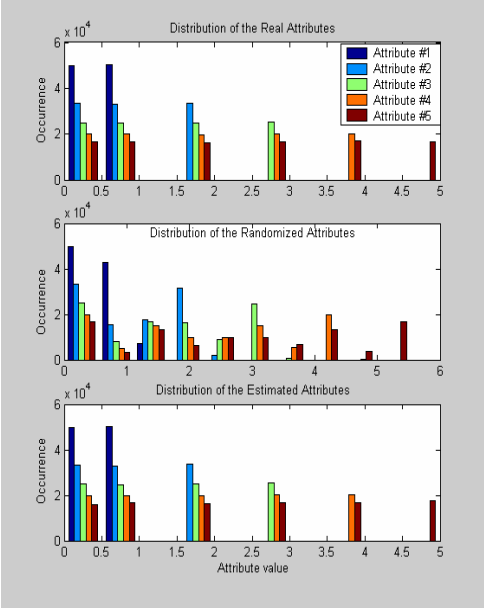

Figure (13) Distribution of original, randomized, and estimated data for attribute to noise ratio "R" = 20dB using "Uniform" Noise

The second step in the process of the Step-class algorithm is to perform the classification and assign a class label for each tuple for which the class label is missing or unknown. To measure the classification accuracy of the algorithm, the strength of the randomizer is changed relative to the attribute value.

Figure (14) shows the classification accuracy of the Step-class algorithm when the randomizer is Gaussian. One can see that the classification accuracy changes from 30% to approximately 100% depending on the value of R. When the ratio (R) is low, (high level of privacy), the classification correct rate is poor (30%) and consequently the classification error is high and the classification accuracy increases as this ratio increases (less privacy), which agrees with the analytical results.

Figure (15) shows the classification accuracy of the Step-class algorithm when the randomizer is Uniform. One can see that when the ratio (R) is low, the classification error is high . The classification error decreases as this ratio increases.

From Figure (14) and Figure (15), it is clear that, when the value of 'R' is small, the classification error is high using either Uniform or Gaussian noise but the classification error when using Gaussian noise is less than when using Uniform noise.

When the value of 'R' goes high, the classification error becomes low for both of them but also it goes to approximately zero a little bit faster than when using Uniform noise.

From the above mentioned results, it is clear that, the classification error is high in case of using Uniform noise than the case of using Gaussian noise. This could be explained as, when using a Uniform noise, addition of uniformly distributed random variable Y to the original attribute; results in a shift of the distribution of the attribute to the right. Consequently, during classification; some real values of the attribute will not appear at all, especially when σ_y^2 is large which leads to a higher rate of classification error.

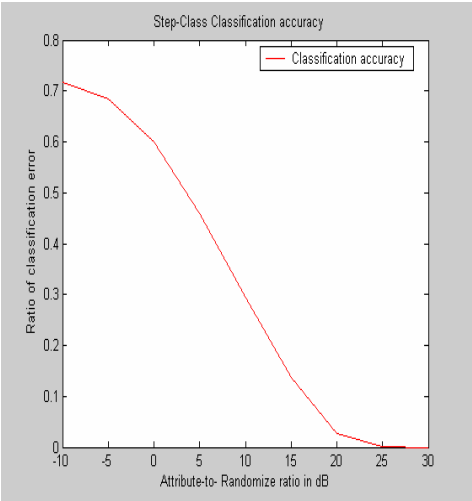

Figure (14) Step-Class classification accuracy using "Gaussian" Noise

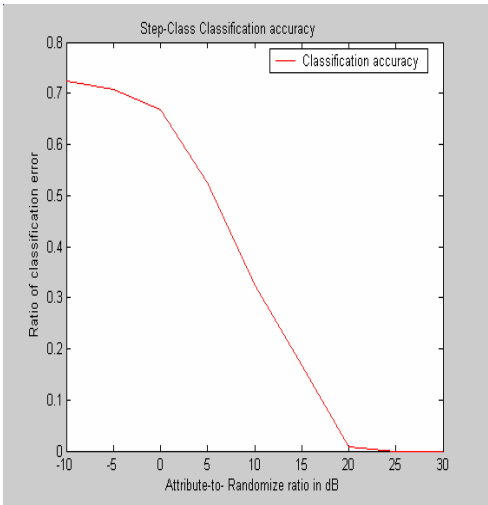

Figure (15) Step-Class classification accuracy using "Uniform" Noise

4.4 Evaluation of the Global-decision algorithm

In this section we present the second algorithm (Global-decision) which is modeled as shown in Figure (16). The model is based on the global decision method described in section 3.3.1 which allows dependencies between attributes to be there and specify joint conditional probability distributions.

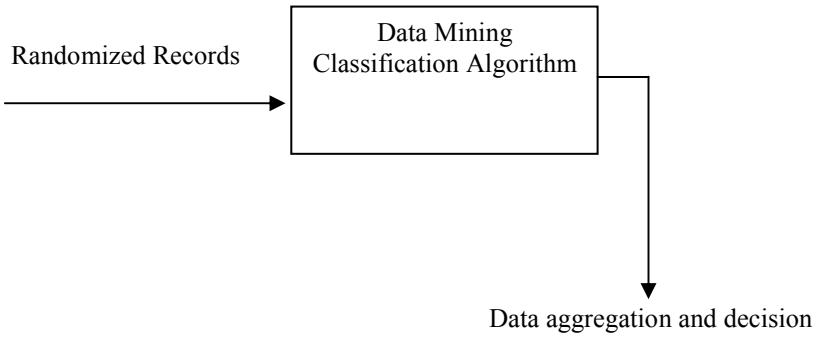

Figure (16) Global-decision Model block diagram

We recall the decision function defined in 3.3.1 as $\widehat{C} = C_k$ *iff* d_k is the minimum distance for k=1,2,…. M
Where d_k is the Euclidian distance between vector X_k and the observed vector $\underline{Z_0}$ where

$$d_k = \sqrt{\sum_{i=1}^{M} (Z_0(i) - X_k(i))}^{2}$$

The input to this algorithm is a set of randomized records and it is assumed that the distribution of the randomizer is known. Those two pieces of information, along with a standard statistical theorem called Bayes' rule, allow the data mining algorithm to estimate the true class of each record with certain degree of accuracy.
In Global-decision method, the classification algorithm is concerned with decision making on the randomized record. We don't make decision for individual attributes; rather we make single decision for the overall randomized record in one step.

In the experimental simulation, each attribute is randomized by a Gaussian and Uniform random variable y_i that is independent from all other attributes. For simplicity, we assume that we have a finite set of classes, and it is required to associate "Classify" each randomized record to one class, such that, if the real record is the one available to the classifier, it will be associated to the same class of the perturbed record.

During the simulation, we have used the same data set with the same values of the ratio between the randomizer strength and the attribute values ranging from (-10db to 30db) which means that the strength of the randomizer starts from 10 times the strength of the attribute value (R = -10db) and decrease until the attribute value becomes 1000 times the randomizer strength.

Figure (17) and Figure (18) show the classification accuracy of the Global-decision algorithm using Gaussian and Uniform randomizer with the change of the ratio R respectively. One can see that, when the ratio (R) is low, the classification error is high and the classification error decreases as this ratio increases. From the

above mentioned results, it is clear also that, with high value of "R", the classification error is a little bit high in case of using Uniform noise than the case of using Gaussian one.

Since all the attributes are positive values, the randomizer is uniformly distributed over positive interval, so the net effect of randomization will be translation of the attribute to the right, but the distribution of the perturbed data will look like the original one whatever the value of the ratio is.
This explains why the result in Figure (17) is a little bit better than the corresponding one shown in Figure (18).

Figure (19) and Figure (20) show the classification accuracy of the two algorithms (Step-class, Global-decision) using Gaussian and Uniform noise respectively. One can see that the Step-class performance is slightly better than the Global-decision in case of small values of 'R', and when 'R' increases, their performance is the same. On the other hand, the Global-decision is better than the Step-class in performing the classification of the perturbed records in one step rather starting by reconstructing the original distribution for each attribute and then performing the classification.

The analysis of the mentioned results is related to the proposed data set used in the simulation. Changing this data set may lead to some deviation in these results.

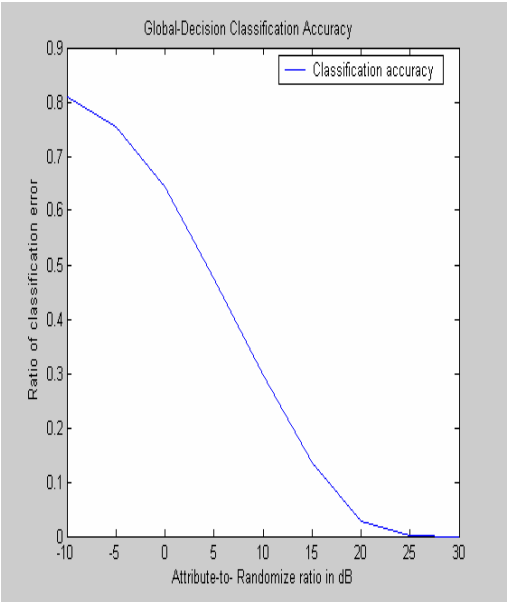

Figure (17) Global-decision classification accuracy "Gaussian" Noise

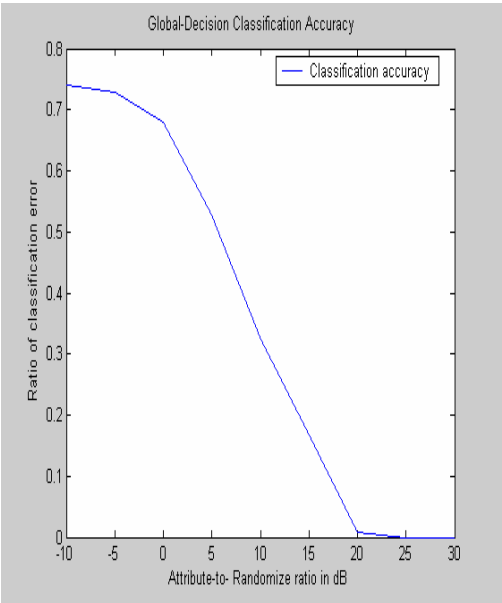

Figure (18) Global-decision classification accuracy "Uniform" Noise

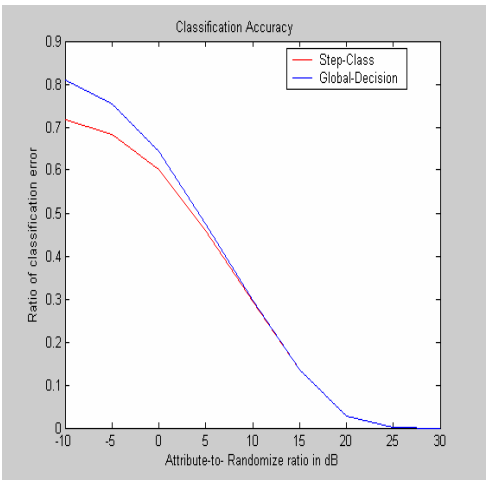

Figure (19) Step-class, Global-decision classification accuracy using "Gaussian" Noise

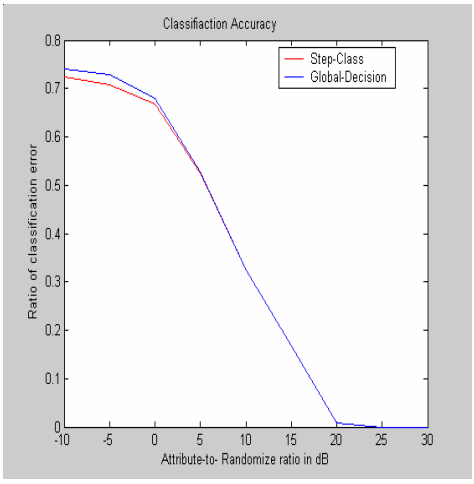

Figure (20) Step-class, Global-decision classification accuracy using “Uniform" Noise

6. Conclusion and discussion

In this paper, the implementations of two new Bayesian classifier algorithms (Step-Class, and Global-Decision) that can be used to classify perturbed data with high accuracy were discussed. These two algorithms were experimented and their performance was checked by measuring their classification accuracy (ratio of correct classification) at different values of the ratio between the strength of the randomizer relative to the average of the set of values that each attribute takes. The results showed that the Step-Class algorithm has better performance results (in terms of classification accuracy) than the Global-Decision algorithm. The results are related to the proposed data set used in the simulation. Changing this data set may lead to some deviation in these results.

REFERENCES

[1] Fahmy A., Fakhry M., Ismail H., El-Zeweidy M. A "A comparative study of algorithms that do data mining keeping data privacy" 11th international conference on Aerospace sciences and Aviation Technology May 17:19, (2005)

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

[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] Jiawei Han, Micheline Kamber , "Data Mining : Concepts and techniques", 2001

[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] Wozen Craft , "Principles of communications" , Artch house 1989.

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

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

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

[14] Chai Wah Wu, "Privacy Preserving Data Mining: a Signal Processing Perspective and a Simple Data Perturbation Protocol.", IBM Research Division. Computer Science, RC22815 (W0306-040) June 9, 2003. Thomas J. Watson