December 2003

Privacy Protection in Data Mining

Jinquan Li
University of Illinois

Michael Shaw
University of Illinois

Fu-ren Lin
National Sun Yat-sen University

PRIVACY PROTECTION IN DATA MINING

Jinquan Li
University of Illinois at Urbana-Champaign
jli4@uiuc.edu

Michael J. Shaw
University of Illinois, Urbana-Champaign
mjshaw@uiuc.edu

Fu-ren Lin
National Sun Yat-sen University
frlin@cc.nsysu.edu.tw

Abstract

Data mining as one of the important means to discover interesting and potential useful patterns or knowledge from large data sources has been widely used for improving business intelligence. Since some data items may be specific to individuals, companies increasingly pay attention to privacy issues while implementing business intelligence solutions. In this paper, we present a framework for privacy-enhancing data mining and develop such privacy-enhancing technologies as attribute selection, discretization, fixed-data perturbation, probability distribution, and randomization. Specifically, we address the issue of privacy protection through using the attribute selection, discretization, and randomization techniques and give an example of inducing the decision-trees from training data in which the values of sensitive attributes have been modified by using these techniques. The results show that we can achieve comparative predictive accuracies without accessing the actual values of the sensitive attributes.

Keywords: Data mining, privacy-enhancing system, information loss, classification, business intelligence

Introduction

Data mining as one of the important means to discover interesting and potential useful patterns or knowledge from large data sources has been widely used for improving business intelligence. Organizations have employed data mining for a variety of services, such as market segmentation, customer profiling, fraud detection, evaluation of retail promotions and credit risk analysis. For example, by mining demographic and socio-economic data, marketers can develop products and promotions to appeal to specific consumer groups, dramatically reducing advertising budgets and boosting revenue. However, since a large amount of information in enterprise databases is often specific to individuals, there has been an increased concern about the privacy of the data in recent times.

Data mining is particularly vulnerable to misuse when there may be some sensitive information that can be extracted by malicious miners if they have access to the micro data provided by a company for mining purposes. Enterprise databases may include variety of data, such as name, address, and telephone number. However, they may also include much more sensitive information about individual's income and financial information, medical, criminal, and education history, property ownership, genetic makeup, insurance files, credit card numbers, welfare bills, employment records, trade secrets. Since a large amount of such data is often specific to individuals, depending on the nature of the data, owners may not be willing to divulge the sensitive values of the micro data. For example, a company surely would not agree to open trade secrets to a competitor, and a person would disclose data on his health status to his doctor, but possibly not to his employer or health insurance companies. Personal privacy is vital to an individual's dignity and free existence while business intellectual property provides an economic advantage to a business relative to its competitors. Failure to protect data privacy in data mining can undermine customer confidence for the company, destroy its competitiveness in the market and expose it to lawsuits. Customers avoid companies that do not protect their credit card numbers or personal information. Therefore, data mining applications demand particularly sensitive treatment to avoid misusing the micro data.

A simple solution to this problem is to eliminate sensitive attributes while preparing input data for data mining tasks. But this may enforce so many restrictions on data, that it would prevent useful mining. Based on considerations of balancing privacy protection and data quality, can we develop new data mining techniques that incorporate privacy concerns? Since the primary task in data mining is the development of models of aggregated data, can we develop accurate models without access to sensitive values in individual data records? This paper concentrates on technological solutions that are able to retain privacy by accessing the information implicit in the original attributes. We address the concrete problem of building decision-tree classifiers and show that it is possible to develop accurate models while respecting individuals' privacy concerns. We review previous research in this field in Section 2. In Section 3, we present a framework for privacy-preserving data mining. In Section 4, we discuss some privacy-enhancing methods. In Section 5, we give an example of inducing the decision-tree classifier from training data in which the values of sensitive attribute values have been modified by using these techniques. We conclude with a summary and directions for future work in Section 6.

Related Work

Confidentiality issues for statistical data have been extensively studied (Adam and Wortman 1989). The proposed approaches can be broadly classified into query restriction and data perturbation. The query-restriction approach provides privacy protection through one of the following means: restricting the size of query result, controlling the overlap among successive queries, keeping audit trail of all answered queries and constantly checking for possible compromise, suppressing data cells of small sizes, and partitioning the population into a number of mutually exclusive clusters. Among these methods, partitioning is closely related to the privacy-preserving techniques discussed in this paper. The idea of partitioning is to partition individual entities of the population into a set of disjoint, mutually exclusive subsets. The statistical properties of these subsets constitute the raw data available to the database users. As long as the subsets do not contain precisely one individual entity, a high level of security can be attained.

The methods based on the data perturbation approach fall into the following three main categories: fixed-data perturbation (Conway and Strip 1976), probability distribution (Liew et al. 1985 and Reiss 1984), and value dissociation (Conway and Strip 1976). The fixed-data-perturbation category is a widely used method for confidentiality protection in statistical databases through adding zero mean noise to the values of the sensitive attributes in the database. The probability distribution category considers the data to be a sample from a given population that has a probability distribution. This method replaces the original database by another sample from the same distribution or by the distribution itself. The two methods identified in this category include data swapping and probability distribution. Data-swapping is a data transformation technique that involves finding transformations that map the original data matrix into a new database which exhibits the same statistics. Since finding a general data swap is thought to be an intractable problems, Reiss (1984) suggested a method that deals with multicatagorical attributes called "approximate data swapping." However, the precision resulting from this method may be considered unacceptable since, as shown in Reiss (1984), the method may in some cases have an error up to 50%. In general, data swapping has not been developed enough to be seriously considered for security control in databases. The probability-distribution method calls for replacing the original database by its (assumed) probability distribution. The method introduces the sampling bias to the results. The bias results from sampling from a population that is not the true-target population. Finally, the value dissociation category represents another alternative for security control. In this method, actual values are dissociated from the actual records in which they occur. Both the probability distribution method and this method are applicable to both categorical and continuous attributes while the fixed-data-perturbation category is applicable only to fields with continuous values.

Adopting some approaches from the statistics literature, data mining researchers recently proposed some innovative approaches for privacy-preserving data mining (Agrawal and Srikant 2000, Agrawal and Aggarwal 2001). Agrawal and Srikant (2000) (we shall henceforth refer to this technique as the AS algorithm) addressed the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level before conducting the mining. They investigated the specific task of building a decision-tree classifier from training data in which the values of individual records have been perturbed. They considered two methods for perturbing values: value-class membership that discretizes the continuous values into discrete values and fixed-data perturbation that adds zero mean noise to the sensitive attribute. The privacy measure was based on how closely the original values of a modified attribute can be estimated. While it was not possible to accurately estimate original values in individual records, they proposed a reconstruction procedure to accurately estimate the distribution of original data values by using Bayes' rule. By using the reconstructed distribution, they were able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. Agrawal and Aggarwal (2001) introduced an Expectation Maximization (EM) algorithm for distribution reconstruction, which is more effective than the AS algorithm in terms of the level of information loss. Given the perturbed values and the noise density function, the EM algorithm is able to approximate the original distribution to a very high

degree of precision. The EM algorithm quantifies privacy based on the differential entropy of a random variable and the concept of mutual information between the original and perturbed records.

While these papers extend the fixed-data perturbation method for data mining, they also preserve the limitations of this method as well. First of all, the most serious limitation is that these methods are applicable only to continuous attributes since arithmetic operations like adding noise are defined only for continuous variables. The database attributes can be of various types including categorical or continuous. Hence ideal are the methods that are applicable to both continuous and categorical attributes. Second, the methods suffer in terms of scale. For example, perturbing a salary of $100,000 by 2000 would be considered a compromise while at the same time perturbing a salary of $10,000 by 2000 would preserve the confidentiality of the data. Several alternatives to this basic method have been suggested in (Cios *et al*. 1998). One alternative is to apply multiplicative rather than additive perturbation, thus overcoming the scale problem. Finally, although the methods look promising, its security aspect needs further investigation. In particular, if the estimated probability-distribution function is a very precise description of the original data, there is hardly any protection against partial disclosure (i.e., the disclosure of the distribution). If partial disclosure is not important, why do we compute the original distribution and give it to the miner in order to save the miner's cost to reconstruct the distribution?

While the two recent papers focused on continuous attributes through data perturbation and reconstruction process, this paper considers both continuous and categorical attributes, and addresses the issue of privacy protection through the privacy-preserving data preprocessing techniques. The database attributes can be of various data types including continuous, categorical and complex data such as documents, graphics, images, audio, and video. Each attribute has a different physical meaning and hence has a different schema. It is desirable to have generic privacy-enhancing methods for the attributes with various data types. Since many of sensitive attributes are categorical and the analysis of continuous attributes can also be based on suitably categorical versions of these attributes through the discretization process, we are motivated to take categorical variables into consideration and develop a framework for privacy-enhancing data mining and some effective technical solutions for preserving privacy in data mining.

A Framework for Privacy-Enhancing Data Mining

In Figure 1, we show a framework for privacy-enhancing data mining. This framework consists of the following components:

Data. The database attributes can be of various data types including numeric, categorical, spatial, text, or image. It is desirable to have a method that can be applied to control the privacy of the attributes of various data types. Generally speaking, we should focus on categorical data models since attributes of other data types can be easily transformed into categorical variables by categorization. The analysis of attribute sensitivity is important in the context of privacy-preserving data mining. A data set may contain both public information and private information. When providing data, individuals would dictate the sensitivity of each attribute.

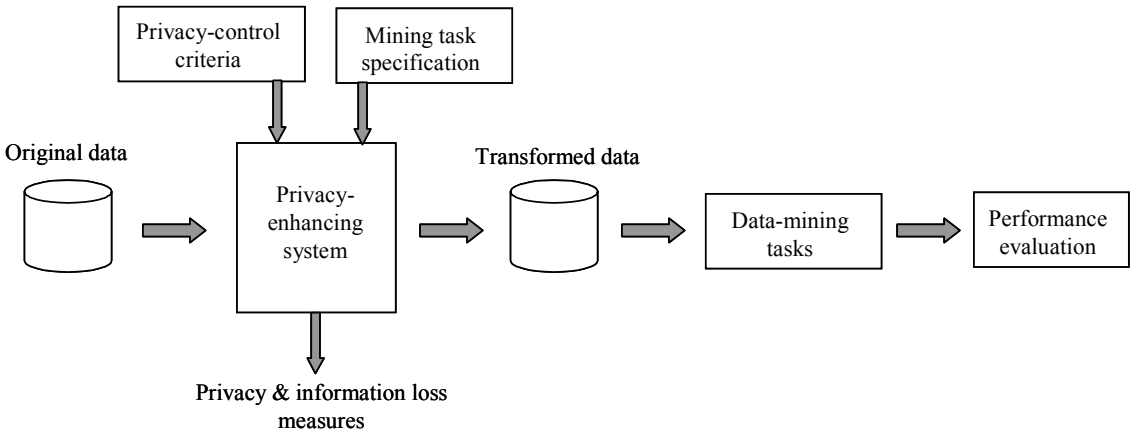

Figure 1. A Framework for Privacy-Enhancing Data Mining

Privacy-control criteria. The privacy metrics used so far (Agrawal and Srikant 2000, Agrawal and Aggarwal 2001) are only good for continuous attributes. It is ideal to have a privacy measure that can be applied to control the privacy of both continuous and categorical variables. Theoretically the privacy metric should be based on how likely the original values of a modified attribute can be identified. In this paper, however, we consider a privacy-preserving method to be acceptable if it prevents complete disclosure and satisfies the given privacy-control threshold. A complete disclosure occurs if a user sees the actual sensitive values in data. The notion of privacy-control threshold is based on the estimated probability that the intruder can identify the original values.

Privacy-enhancing methods. The performance of the privacy-enhancing methods is mainly judged on two factors: one, the precision of the mined results should be as high as possible; second, the methods should preserve privacy. The mining algorithms used influence the quality of the mined results, but the most influencing factor is the data used. Ideally, we should also provide data miners with all the relevant data required for the mining purposes. "Garbage in, garbage out." Data quality directly determines the usefulness and accuracy of the mining results. But the privacy-enhancing methods may enforce some restrictions on data. An ideal method is able to retain privacy while accessing the information contained in the original databases as much as possible. In addition to these two criteria, the methods are aimed at improving the efficiency of the mining process and the quality of the results. For example, too many attributes could slow down the mining process considerably and make the mining results hard to understand. It might be a good idea to remove the attributes irrelevant to the mining task during data preprocessing. An ideal method should also become a more automated process that helps companies easily to implement privacy protection.

Data mining tasks. We may group the mining tasks under three primary paradigms: predictive modeling, discovery, and deviation detection. The goal of predictive modeling is to find patterns involving attributes for predicting or classifying the future behavior of some entity. Predictive modeling applications make classifications and predictions. Discovery applications are exploratory approaches for data analysis. They employ techniques that analyze large data set to find association rules (or patterns), or to find clusters of examples that can be grouped together. Deviation detection performs automatic anomaly detection. The system first identifies what is usual and establishes a set of norms through pattern discovery. The examples that deviate from the norm are then identified as unusual. Each mining task analyzes data from a different angle and produces a different mining output. Therefore, they may need different information contained in the data and require different privacy-preserving methods that access the information.

Privacy-Enhancing Data Preprocessing Methods

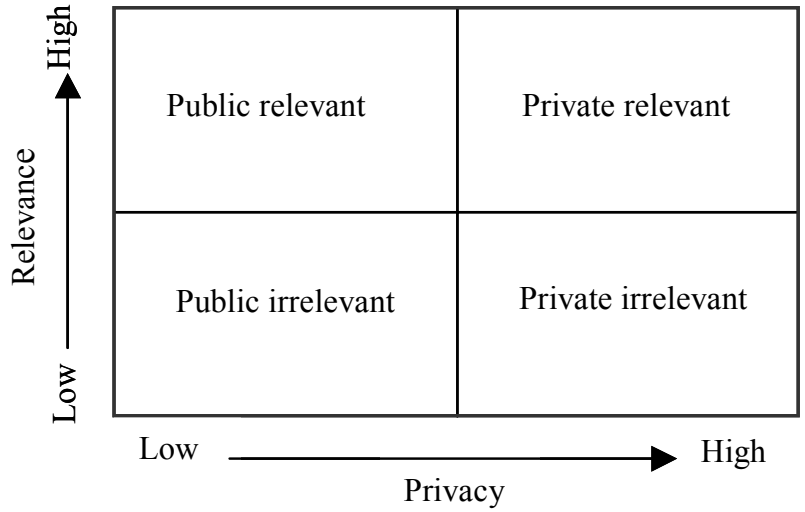

Figure 2. Contingence Table of Attribute Sensitivity and Relevance

There are three main steps in the data mining process: (1) data preprocessing where the data is prepared for the mining purposes; (2) processing where algorithms are used to process the data in order to find hidden knowledge; (3) analysis where the output is

evaluated to determine how relevant the findings are. Obviously, we should address our privacy concerns during the preprocessing step. Our argument is simple: before we make information available, we must make sure that we are not giving out more than we should.

In this paper, a database or data set D is modeled as an $n \times m$ matrix, where each of n rows represents an individual record or example, and each of the m columns represents an attribute. We denote a set R of records represented by a set F of attributes in the database by $A_1, A_2, \ldots, A_m$. The data elements are the values of the matrix and correspond to the values for the given record and the given variable. The techniques that can prevent disclosure of sensitive values in the database fall into two categories: data reduction, restricting the information to be released, and data perturbation, modifying the original database before it is released. Each of them is composed of a number of techniques.

Privacy-Enhancing Data Reduction Techniques

A large database is now measured gigabytes and terabytes and its analysis is becoming very costly and inefficiently. The complexity when dealing with the databases has led researchers to develop data reduction techniques in order to reduce the cost of data mining. Interestingly, these techniques are also very useful for privacy protection.

Attribute selection. Attribute selection is very efficient when we preserve privacy from considerable attributes in very large databases. For example, if the task is to predict whether or not they are likely to purchase a new music CD at a store, attributes such as the customer's social security number and telephone number are likely to be irrelevant, unlike attributes such as age or music taste. Two aspects of the method are attribute relevance analysis and attribute sensitivity analysis (Figure 2). The idea behind attribute relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class or concept. Such measures include information gain, the Gini index, uncertainty, and correlation coefficients. If information gain is used as a measure, for instance, the attribute with the highest information gain is considered the most relevant attribute of the given data set. By comparing the information gain for each attribute, we therefore obtain a ranking of the attributes. The ranking and the specified threshold can help us to select the attributes to be used in data mining. The goal is to preserve general statistics based on attributes that have a high impact on the mined results after removing irrelevant attributes in order to improve the efficiency of the mining process.

In addition to attribute relevance analysis, attribute privacy analysis is also important. A data set may contain both public information and private information. Figure 2 shows the contingence between privacy and relevance. The idea behind attribute sensitivity analysis is to ensure that all public attributes are totally accurate while sensitive attributes are protected. Obviously, the attributes irrelevant to the mining task or redundant should be removed. One common privacy-preserving technique is to anonymize information by removing identifiers such as names, addresses, telephone numbers, social security numbers, drivers-license numbers, product names and product numbers etc. All these identifiers are very sensitive but usually have very little discriminatory power. After the removal of the private irrelevant attributes, the remaining problem is how to deal with the private relevant attributes. Since the elimination of these variables seriously damages the information contained in the original data, these attributes should be kept in data mining. Fortunately, some of other privacy-enhancing techniques can be used to protect these attributes while accessing the information implicit in these attributes.

By eliminating the irrelevant private attributes, attribute selection may result in a less sensitive input for the mining purposes. Since many privacy concerns involve whether the information is used in an identifiable way, attribute selection is a natural way to address these concerns. This method is also very beneficial for both the mining process and the results. Data with too many attributes could slow down the data mining process considerably. Attribute selection not only speeds up the mining process but also solves the problems associated with the performance and capacity limits of discovery systems.

Discretization. Data usually comes in a mix of continuous and categorical attributes. While the number of continuous values for an attribute can be infinitely many, the number of discrete values is often few or finite. Many studies show that a continuous attribute with very large number of values has negative effects on the mined results and that induction tasks can benefit from discretization: the process to discretize continuous attributes before or during a data mining or machine learning task. The basic idea of discretization is to divide the range of a sensitive attribute into a number of mutually exclusive intervals and interval labels can then be used to replace actual data values. For example, salary may be discretized into 10K intervals for lower values and 50K intervals for higher values. Instead of a true attribute value, the user provides the interval in which the value lies. Discretization is the generalization method used most often for hiding individual values. As long as intervals do not contain precisely one

sensitive value, discretization provides a certain degree of protection. The method is applicable not only to continuous attributes but also to categorical attributes with large number of values.

Discretization helps us to limit access to data, but we still face the following problems. One is about the privacy level. We can control the risk of disclosure by controlling the group size. In other words, the range of the sensitive attribute is partitioned to groups of k or more individuals. The larger the group size, the less likely the individual information can be identified. On the other hand, larger groups may lead to more information loss since the attribute may lose its discriminating power when the values are collapsed. Another issue is the choice of discretization methods. Discretization is a process of finding cut-points and partitioning the values of the sensitive attribute according to the cut-points. The key to discretization is to find the best cut-points and to maximize the information gain of the attribute. Since the entropy (MDLP) method is the best discretization method in supervised learning (Liu *et. al,* 2002), we use the method for the discretization process.

Privacy-Enhancing Perturbation Methods

The perturbation approach can be formulated as follows. Consider a database D and a set S of sensitive attributes. For each sensitive attribute in S, generate the perturbed values for the attribute from the current ones by perturbation, and replace the original database values for the attribute by the perturbed values. Three common perturbation functions are fixed-data perturbation, probability distribution and randomization.

Fixed-data perturbation. A widely used method for privacy protection is to add zero mean noise to sensitive attribute values. Consider a private attribute A_i that is protected by adding a noise variable R, where R is a zero mean random variable drawn from some distribution. A miner attempting to obtain the attribute value for a specific data subject is provided a masked value $M = A_i + U$. Under most additive noise data masking procedures, the noise has a zero mean distribution. For example, U has a normal distribution with mean $\mu = 0$ and standard deviation σ. This perturbation technique is robust (resistant to users' supplementary knowledge), but affecting the quality of data available for statistical analysis since the method has a large risk of introducing bias to the mining results. As the size of the database increases, the bias becomes smaller but at the expense of allowing partial disclosure of sensitive attributes. The method has an advantage over the probability-distribution method because the way in which noise is added to the data is much clear and therefore better suited for statistical analysis. As indicated above, the method also suffers in term of scale. But the most serious limitation is that the method is applicable only to continuous values.

Probability distribution. The probability distribution method calls for replacing the original database by its probability distribution. It consists of three steps:

(1) identify the underlying density function of the attribute values and estimate the parameters of this function,
(2) generate a sample series of data from the estimated density function of the sensitive attribute. The new sample should be the same size as that of the database, and
(3) substitute the generated data of the sensitive attribute for the original data in the same rank order. That is, the smallest value of the new sample should replace the smallest value in the original data, and so on.

This method is equivalent to the data perturbation method. Hence it introduces the sampling bias to the results. The bias results from sampling from a population that is not the true-target population. For a data set of small size, the noise introduced by this method is larger. But as the size of the database increases, the bias becomes smaller but less security of sensitive attributes is achieved. Partial disclosure is easily possible since the noise added to the sensitive attribute becomes rapidly small.

Randomization. We propose the randomization methods for privacy protection in data mining. Consider a sensitive attribute X that is protected by the randomization techniques. The idea of randomization is to randomly shuffle the values of X between records in order to dissociate the attribute value from the actual record it represents. The randomization process basically applies a sequence of elementary swaps to the original values of X. An elementary swap consists of two actions: a random selection of two records i and j; an interchange of the values of X being swapped, for records i and j. A set P of pairs (i, j) is then determined, where i and j are distinct rows of the set $\{1, \ldots, n\}$ of the data set. We consider two randomization algorithms that differ in the determination of the set P:

- Global randomization: Randomly shuffle all the values of X between records and construct perturbed data set by using randomized values of the attribute. The set P includes all the pairs of distinct rows of the data set.

- Randomization by class: First split the values of X by class, then for each class, randomly shuffle the values of X within the class. Construct perturbed data set by using randomized values of the attribute. The determination of the set P is constrained within the class.

Global randomization provides a high level of privacy because the relation between X and the class attribute C is totally perturbed. However, it affects the joint distribution between X and C. For global randomization, the associations between the variables X and C are likely to be attenuated. Therefore, global randomization may seriously damage the information gain of the attribute X and cause a high level of information loss.

On the other hand, randomization by class works as follows. Suppose that each record is partitioned into three parts, X, Y, and C, where $Y \in F - \{X, C\}$ is the set of attributes excluding the sensitive attribute X and the class attribute C. C defines a categorical variable with class labels. With randomization by class, the set P include only pairs of records for which C is constrained to be the same, that is $C_i = C_j$ if $(i, j) \in P$. In this case the joint distribution of the pairs of values (C_i , X_i) and (C_i, Y_i) is preserved by swapping. Since this method preserves the joint distribution of the variables X and C, the associations between the variables X and C are maintained and the information loss is minimized. The method is also likely to provide a high level of privacy because the number of classes is usually small.

The randomization techniques have some significant advantages over fixed-data perturbation, such as

(1) the methods preserve privacy in large databases and are applicable to the database attributes of various data types including continuous, categorical, text, spatial, or image;
(2) the distribution of the values of the sensitive attribute is naturally preserved since the perturbed values of the attribute are actual values in the original database;
(3) since actual values are returned, schema of those values is preserved. For instance, if all salaries are in even thousands of dollars, this fact would be preserved under the methods.

The experimental results in Section 5 show that it is possible to transform the database by randomly shuffling the values of private attributes so that the models produced by the decision trees on the modified data set correspond to the models produced from the original database.

Inducing the Decision-Trees over Transformed Data

Classification is probably the most common data mining activity today. This task takes as input a set of classes and a training set consisting of pre-classified cases, and builds a model of some kind that can be applied to unclassified data in order to assign it a class. Decision-tree classifiers are relatively fast, yield comprehensible models, and obtain similar and sometimes better accuracy than other classification methods. In this section, we want to evaluate the effect of the privacy-enhancing methods on classification. For this purpose, we have built a proof-of-concept privacy-enhancing preprocessing system. The system is written in C++ and implements a variety of functionalities, such as feature selection, entropy-based discretization, user-defined discretization, global randomization, randomization by class, etc. The inputs are the original data set and the privacy specification. The outputs are the transformed data set. We investigate if we are able to build classifiers with the modified training set whose accuracy is comparable to the accuracy of classifiers built with original data. Thus, we need a classification algorithm. C4.5 (Quinlan, 1993) is chosen for experiments because it can handle both data types, i.e., continuous as well as categorical, and it is conveniently available and widely used so that a reader can easily repeat the experiments here. Furthermore, C4.5 has become a de facto standard for comparison in machine learning.

Table 1. Mutual Information for Each Attribute

Attributes	Information Gain
Number of times pregnant	0.061825
Plasma glucose concentration	0.304201
Diastolic blood pressure	0.05931
Triceps skin fold thickness	0.081664
2-hour serum insulin	0.232564
Body mass index	0.123164
Diabetes pedigree function	0.010896
Age	0.140941

Table 2. Results for the Pima Indians Diabetes Data

	Accuracy on training data	Accuracy on test data
Original Data	88.60%	74.10%
Removal of *number of times pregnant*	85.50%	74.10%
Experiment 1	81.10%	75.30%
Experiment 2	81.40%	74.50%
Experiment 3	88.20%	69.10%
Experiment 4	87.60%	71.20%

The Pima Indians Diabetes data is selected from the UC Irvine machine learning data repository (Merz and Murphy, 1996). This dataset contains information on patients of Diabetes. All patients are females at least 21 years old of Pima Indian heritage. The data is interesting because there is a good mix of attributes – continuous, categorical with small numbers of values, and categorical with large number of values. The data consists of 8 variables plus the class attribute. The data set contains 768 cases. Our program randomly split the data into a training set of 576 training cases and a test set of 192 cases. Among these attributes, the attributes *Number of times pregnant* and *Age* are supposed to be sensitive. A decision-tree classifier is built from the training data by C4.5. The predictive accuracies on the original training data and test data are 88.6% and 74.1% respectively.

We use information gain to quantify the relevance of an attribute with respect to the class attribute. We run SFG (Liu and Setiono 1998), an application for sequential forward feature selection using information gain, on the original data. Table 1 reports the information gain for each attribute. Given a threshold value of 0.1, the sensitive attribute *Number of times pregnant* is an irrelevant attribute but the attribute *age* is a weakly relevant attribute. Therefore, we should eliminate the attribute *Number of times pregnant* without losing much information. The attribute *age* is a weak relevant attribute with relatively large number of values. We consider different privacy-enhancing methods to protect the attribute.

After the removal of the attribute *Number of times pregnant*, we conducted four experiments. In Experiment 1, we eliminate the attribute *Age* from both the training and test sets. We use C4.5 to build the decision tree for the new training set with the remaining attributes. The new test set is used to evaluate the classification accuracy. In Experiment 2, we run the entropy-based (MDLP) discretization method on the attribute *age* with a minimum group size of 5 to get cut-points, use the cut-points to discretize the attribute *age* for both the training and test set. The minimum group size of 5 enables us to prevent the complete disclosure and to control the disclosure risk effectively. We use the discretized training set to build the decision tree and use the discretized test set to evaluate the classification accuracy. In Experiment 3, the randomization by class method is used to perturb the attribute *Age* in both the training and test data while other attributes are unchanged. C4.5 is used to build the decision tree for the transformed training data and the transformed test set is used to evaluate the classification accuracy. In Experiment 4, the Global randomization method is used to modify the attribute *Age* while other attributes are unchanged. Similarly, we use C4.5 to measure the predictive accuracies for both the transformed training and test data. Table 2 reports the results of these experiments. As shown in Table 2, the best privacy-enhancing method for the attribute *age* is to eliminate or discretize it. This fact can be explained as follows. The attribute *age* is a very weakly relevant attribute and has very little impact on the discovery of decision trees. The best strategy is to remove or discretize it. Otherwise, its large number of values has negative effectives on the classification accuracies. The randomization techniques are unable to reduce the number of the values of the attribute and consequently yield low predictive accuracies. Discretization can lead to improved accuracies since it decrease the number of the values of the attribute. From the

above results, we can conclude an important observation: the application of the privacy-enhancing data mining methods may preserve the privacy with minimum information loss, which causes the low predictive accuracies.

Conclusions and Future Work

In this paper, we develop a framework for privacy-enhancing data mining and effective technical solutions for preserving privacy in data mining. The techniques can be used to protect the privacy of individual values with little negative effects on the precision of the mined results. We address privacy concerns by applying privacy-enhancing transformation to the original data during the data preprocessing step and gave an example of inducing the decision-tree classifier from training data in which the values of sensitive attribute values have been modified. The experimental results show that we can achieve comparable prediction accuracy from the transformed data without accessing sensitive attributes.

The paper only considers inducing the decision-tree classifier from the data sets generated by different privacy-enhancing techniques. Further research about privacy protection should try on other data mining tasks, such as association rule mining, clustering, deviation analysis etc. We also notice that preserving just the sensitive information is not sufficient in many cases because the unauthorized data miner may find the inference paths from the public to the sensitive information. In this case, link analysis is necessary to preserve privacy in data mining.

References

Agrawal, D. and C.C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proceedings of the Twenty ACM SIGMOD_SIGACT-SIGART Symposium on Principles of Database Systems, 2001, pp. 247-255.

Agrawal, R. and R. Srikant. Privacy preserving data mining. In Proceedings of the ACM SIGMOD, 2000, pp. 439-450.

Adam, N.R. and J.C. Wortman. Security-control methods for statistical databases. ACM Computing Surveys (21:4), 1989, pp. 515-556.

Cios, K.J., W. Pedrycz and R.W. Swiniarski. Data Mining: Methods for Knowledge Discovery, Massachusetts: Kluwer Academic Publishers, 1998.

Conway, R. and D. Strip. Selective partial access to a database. In Proceedings of ACM Annual Conference, 1976, pp. 85-89.

Liew, C.K., U.J. Choi, and C.J. Liew. A data distortion by probability distribution. ACM TODS (10:3), 1985, pp.395-411.

Liu, H., F. Hussain, C.L. Tan and M. Dash. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery (6), 2002, pp. 393-423.

Liu, H. and R. Setiono. Feature Extraction, Construction and Selection: A Data Mining Perspective: Kluwer International Series, 1998.

Merz, C.J. and Murphy, P. UCI repository of machine learning databases. **http://www.ics.uci.edu/~mlearn/MLRepository.html**. Irvine, CA: University of California, Department of Information and Computer Science, 1996.

Quinlan, J.R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

Reiss, S.P. Practical data-swapping: the first step. ACM TODS (9:1), 1984, pp. 20-37.