ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/220373844

Protecting Data through Perturbation Techniques: The Impact on Knowledge Discovery in Databases.

Article · January 2003

DOI: 10.4018/978-1-59140-471-2.ch003 · Source: DBLP

CITATIONS
19

READS
2,967

2 authors:

Rick L Wilson
Oklahoma State University - Stillwater
59 PUBLICATIONS **1,833** CITATIONS

SEE PROFILE

Peter A. Rosen
Pittsburg State University
24 PUBLICATIONS **767** CITATIONS

SEE PROFILE

All content following this page was uploaded by Rick L Wilson on 19 June 2014.

The user has requested enhancement of the downloaded file.

Protecting Data through 'Perturbation' Techniques: The Impact on Knowledge Discovery in Databases

Rick L. Wilson and Peter A. Rosen, Oklahoma State University, USA

ABSTRACT

Data perturbation is a data security technique that adds 'noise' to databases allowing individual record confidentiality. This technique allows users to ascertain key summary information about the data that is not distorted and does not lead to a security breach. Four bias types have been proposed which assess the effectiveness of such techniques. However, these biases only deal with simple aggregate concepts (averages, etc.) found in the database. To compete in today's business environment, it is critical that organizations utilize data mining approaches to discover additional knowledge about themselves 'hidden' in their databases. Thus, Database Administrators are faced with competing objectives: protection of confidential data versus data disclosure for data mining applications. This paper empirically explores whether data protection provided by perturbation techniques adds a so-called Data Mining Bias to the database. The results find initial support for the existence of this bias.

Keywords: Data Perturbation, Data Mining, Data Security, Confidentiality

INTRODUCTION

Today, massive amounts of data are collected by organizations about customers, competitors, supply chain partners and internal processes. Organizations struggle to take full advantage of their data, and discovering 'unknown' bits of knowledge in their massive data **stores remain**s a highly sought after goal.

Database and data security administrators face a problematic balancing act regarding access to organizational data. Sophisticated organizations that do take advantage of data mining and knowledge discovery algorithms (e.g., inductive learning algorithms, neural networks, etc.) to find previously unknown 'patterns' in their data benefit greatly by having access to large data stores of individual records.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

However, the need to protect individual 'confidential' data elements in an organizational database from improper disclosure is another important issue faced by the database administrator. This protection concerns not only traditional data access issues (i.e., hackers and illegal entry), but also masking individual confidential record attributes to inhibit individual record identification even by authorized users.

Techniques that seek to accomplish masking of individual confidential data elements while maintaining underlying aggregate relationships of the database are called **data perturbation techniques**. These techniques modify actual data values to 'hide' specific confidential individual record information.

Recent research has analyzed increasingly sophisticated data perturbation techniques on the two dimensions of ability to protect confidential data AND (at the same time) preserving simple statistical relationships in a database (means, variances, etc.). However, value-adding knowledge discovery and data mining techniques find relationships that are much more complex than simple averages (such as creating a decision tree for classifying customers, etc.). To our knowledge, no previous studies have explored the impact of data perturbation techniques on the ability for knowledge discovery techniques to work well in a 'perturbed' database. This is the objective of our study.

Specifically, the study empirically examined if classification accuracy of a representative data mining/knowledge discovery tool (QUEST, Low and Shih, 1997) is impacted by various data protection schemes using two 'classic' classification databases, the IRIS and LIVER datasets (Merz and Murphy, 1996). Two data perturbation techniques (one sophisticated, one naïve) will be used to 'hide' confidential attributes of the databases. The classification accuracy of QUEST on the original data and perturbed data will be compared, which will provide insight into potential impacts of this data protection method on data mining approaches.

REVIEW OF RELEVANT LITERATURE

Data Protection through Perturbation Techniques

Organizations store large amounts of data, and most may be considered confidential. Thus, security and protection of the data is a concern. This concern applies not just to those who are trying to access the data illegally, but to those who should have legitimate access to the data.

Our interest in this area relates to restricting access of confidential database attributes to legitimate organizational users (i.e., data protection). **Data perturbation techniques** are statistically based methods that seek to protect confidential data by adding random noise to confidential, numerical attributes, thereby protecting the original data. Note that these techniques are not encryption techniques, where the data is first modified, then (typically) transmitted, and then received, 'decrypted' back to the original data.

The intent of data perturbation techniques is to allow legitimate users the ability to access important aggregate statistics (such as mean, correlations, etc.) from the entire database while 'protecting' the individual identify of a record. For instance, in a simplified case of sales data, a legitimate system user may not be able to access what a particular individual purchased from a store on a given day, but that same user could determine the total sales volume for the store on the same day.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

A group of authors recently examined previously proposed data perturbation methods, and analyzed their effectiveness on various **bias** measures (Muralidhar, Parsa, and Sarathy, 1999). A data perturbation method exhibits bias when the results of a database query on perturbed (i.e., protected) data produces a significantly different 'result' than the same query executed on the original data. Four types of biases were identified, termed Type A, Type B, Type C, and Type D (Muralidhar, Parsa, Sarathy, 1999).

Type A bias occurs when the perturbation of a given attribute causes summary measures (i.e., mean value) of that individual attribute to change due to a change in variance. Type B bias occurs when perturbation changes the relationships (e.g., correlations) between confidential attributes. Type C bias occurs when perturbation changes the relationship (again, e.g., correlations) between confidential and non-confidential attributes. Type D bias deals with the underlying distribution of the data in a database, specifically whether or not the data has a multivariate normal distribution.

It was shown that past methods suffered from one or more of the four aforementioned biases, and thus were inadequate data perturbation techniques (Muralidhar, Parsa, Sarathy, 1999). As an example, consider the most naïve data perturbation method, Simple Additive Data Perturbation (SADP) (Kim, 1986). This approach involves perturbing confidential attributes by adding a noise term with a mean of 0 to the original data. Each confidential attribute in the database is perturbed independently of the other attributes. It can be shown that the method suffers from Type A, Type B, and Type C bias, and thus is inadequate.

Other proposed data perturbation methods include the Correlated-Noise Additive Data Perturbation (CADP) method (Muralidhar, Batra, and Kirs, 1995; Kim, 1986) and the Bias-Corrected Correlated-Noise Additive Data Perturbation (BCADP) method (Kim, 1986; Tendick and Matloff, 1994). These methods were improvements over SADP, but still suffered from biases (CADP—Type A and C, BCADP—Type C).

Multiplicative Data Perturbation (MDP) methods have been proposed as well (Muralidhar, Batra, and Kirs, 1995). Unfortunately, this family of perturbation techniques suffers from all four bias Types.

The General Additive Data Perturbation (GADP) method was proposed (Muralidhar, Parsa, and Sarathy, 1999) as a further improvement to these methods. GADP was shown to possess none of the four biases and is perhaps the 'gold standard' in data protection via perturbation. The process is explained in the following paragraphs.

In a database, we identify the confidential attributes that we would like 'hidden' (even from authorized users) and call this set **X**. All other attributes would be considered the non-confidential attributes (set **S**). A database **U** has i instances with a total set of attributes **X**+**S**. The GADP process will create a perturbed database **P** (based on **U**) that also has i instances with attributes **X**+**S**.

For all attributes in **S**, the attribute value for instance i in database **P** will be the same value of instance i in database **U**. Thus, GADP does not 'hide' (i.e., does not change or perturb) non-confidential attributes. However, for all attributes in **X**, the attribute value for instance i in database **P** will be perturbed (modified) from the value in the corresponding instance i in database **U.**

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

The perturbation process is based upon the original statistical relationships of database **U.** These relationships include the mean values for attributes **X**, the measures of covariance between attribute sets **X** and **S** (i.e., a measure of how the two sets of attributes are related), and the canonical correlation between the attribute sets **X** and **S** (i.e., how well can the actual values of attribute set **X** be predicted by knowing the actual values of attribute set **S**?). Given these statistical properties of **U**, a multivariate normal distribution function is constructed for each instance i (Graybill, 1976). Then, a multivariate random number generator generates the new **X** attribute values for the ith entry in the perturbed database P. This is repeated for all i instances. The use of the three statistical relationships mentioned above AND the actual attribute values from **U** in constructing the multivariate normal random distribution function ensures that all four biases are suppressed in the GADP process. Appendix A provides a more extensive mathematical description of the GADP process presented for the interested reader.

Proposing the Existence of Type 'DM' Bias

The GADP method can be shown to 'protect' confidential attributes appropriately and be theoretically 'bias-free'. However, the four biases previously mentioned represent a very limited view of the value-added capability of a database. Knowledge discovery techniques (i.e., data mining techniques) can identify underlying patterns in a database, often in decision tree or 'rule' form, providing an organization deeper insight (knowledge) about that database. The biases discussed in (Muralidhar, Parsa, Sarathy, 1999) focus only on simple parametric aggregate measures and relationships (means, variances, covariances, etc.).

This study hypothesizes that a 'deeper,' knowledge-related bias may be incurred through these perturbation-based data protection techniques. This bias would measure the alteration or loss of important, knowledge-based relationships. We refer to this as Type Data Mining (DM) bias.

Many data mining approaches useful in knowledge discovery when looking at classification problems have been developed and analyzed (e.g., Lim, Low and Shih, 2000). To assess whether evidence of a Type DM bias exists, this study will use a representative decision tree approach in our experimental manipulations. QUEST (Quick, Unbiased, Efficient Statistical Tree) was chosen due to its accurate performance in a recent study comparing thirty-three knowledge discovery classification tools (Lim, Low, and Shih, 2000).

METHODOLOGY

Databases

If Type DM Bias exists, data mining tools will perform less accurately on perturbed data than they would on the original data set. Specifically in this study, we are investigating the impact of perturbed data in a classification decision task (where each instance of the database is a member of a 'class') within an organizational database.

Two frequently studied data sets in the data mining literature from the UCI Machine Learning Repository were used as surrogates for organizational databases (Merz and Murphy, 1996). These databases have a categorical dependent 'class' variable associated with each database instance, and thus are applicable to be analyzed by a data mining approach. Additionally, they have some other desirable characteristics that are discussed below.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

The first data set was the IRIS Plant Database, chosen for the almost perfect linear separation of its class groups. This data set consists of 150 observations, four numerical independent variables (sepal length, sepal width, petal length, petal width) and the dependent (class) variable of iris plant type (*Iris setosa*, *Iris versicolour*, and *Iris virginica*). The data set is equally distributed among the three classes.

The second data set utilized in the study was the BUPA Liver Disorders Database (Merz and Murphy, 1996). This data set consists of 345 observations, six numerical independent variables (mcv—mean corpuscular volume, alkphos—alkaline phosphotase, sgpt—alamine aminotransferase, sgot—aspartate aminotransferase, gammagt—gamma-glutamyl transpeptidase and drinks—number of half-pint equivalents of alcoholic beverages consumed per day), and a dependent class variable indicating presence or absence of a liver disorder. Two hundred of the observations were classified as Group 1 (58%), and 145 of the observations were as Group 2 (42%). This data set was selected due to the high error rate that other researchers have previously found in classification studies (e.g., Lim, Low, and Shih, 2000). Thus, it is a complimentary example to the 'easy-to-classify' IRIS data set.

Exploring Performance: Creating Perturbed Versions of Databases

To explore how the performance of a data mining/knowledge discovery tool is impacted by a perturbed database, four different 'copies' (i.e., treatment groups) of the two databases were constructed. The first copy is simply the original data (termed ORIGINAL). Obviously, the performance of the data mining tool in predicting class membership on the original data set represents a 'benchmark' performance.

The three other copies of the database were data that had been perturbed using one of three methods. The SADP group applied the naïve, bias-ridden SADP perturbation process to the data. For each confidential attribute, an error term with a normal distribution of mean 0 and standard deviation equal to the standard deviation of the original attribute was created. The perturbed value of database instance was thus the original value plus the corresponding random error term. The SADP perturbed database was included in this study since it is a poor perturbation technique, and should likely serve as the 'lower bound' for data mining tool performance.

The GADP method was used to create the final two perturbed database copies, one that did NOT include the class variable in the perturbation procedure (called GADP) and one that did include it as a non-confidential variable (GADP-W). In an organizational setting, it is unclear whether the classification group of the individual record would be a part of the database and/or known to the user.

Key to the perturbation processes of SADP and GADP is the identification of confidential and non-confidential data. Since we are using two datasets as surrogates of organizational data, we had to choose a strategy for selecting which of the attributes would be confidential. In each dataset (IRIS and LIVER), we chose the n-1 independent variables that had the highest correlation to the dependent class variable as confidential attributes, with the remaining attribute and class variable (when applicable) as non-confidential. This strategy was meant to make the prediction of class as 'hard' as possible for the data mining tool (i.e., the best predictor attributes were the ones being altered in the perturbed data sets). Thus, results could be moderated by this strategy.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

The implementation of GADP required numerous tools. First, a multivariate normal distribution random number generator was necessary to implement the perturbation procedure. The EXCEL add-in NtRand (Numerical Technologies Corporation, 2001) was used to generate the perturbed data according to the GADP methodology outlined in Appendix A. NtRand is based upon significant advances in psuedo-random number generation techniques. The suggested guidelines stated in Matsumoto and Nishimura (1998) and at the Numerical Technologies Corporation web page (2001) were followed such that the differences between the original and perturbed covariance matrices of the data set were minimized. The expected value and variance vectors described in Appendix A were the inputs into NtRand necessary to generate the perturbed data set copies. A VBA application was written which created the GADP and GADP-W datasets using pertinent EXCEL add-ins.

Evaluating Performance: The data mining tool implementation and measurement

The data-mining tool used for this study, the QUEST method (Kim, 1986), is implemented in SPSS's AnswerTree software package. QUEST generates a standard decision tree whose branches can be 'traced' to classify a particular case. The decision tree can also be expressed as a simple set of IF-THEN rules. QUEST default parameter settings were used except for the specification of a stopping rule that limited tree growth to a depth no greater than five levels.

For each different combination of data set (IRIS and LIVER) and perturbation technique (ORIGINAL, SADP, GADP, GADP-W), ten-fold cross-validation was used to determine a robust measure of classification accuracy for the QUEST tool. This is a statistically sound way to determine an accurate measure of classification accuracy (Weiss and Kulikowski, 1991). Cross-validation involves splitting a data set into ten equal (or as equal as possible) parts. Nine parts are used in the training of the data-mining tool (i.e. the original construction of the decision tree), and the remaining part is used to test the ability of the tool to predict unseen cases. Thus, the data-mining tool analyzed each data set ten different times, so that each of the ten parts of the data set was used once as a testing data set. For example, parts 1-9 would be used for training (building the decision tree), and part 10 for testing during the first run of the data-mining tool.

The ten partitions of the data were also stratified. The IRIS data set was split into ten parts of 15 observations each, with each of the ten parts containing five *Iris setosa*, five *Iris versicolour*, and five *Iris virginica* observations. The LIVER data set was split in a similar manner, with 34-35 observations in each part—20 Group 1's and 14-15 Group 2's per part.

Therefore, each data set was 'mined' by QUEST to determine rules of classifications. An instance was labeled 'correctly classified' when the decision tree outcome on the perturbed data set matched the actual class value of the database instance. The correct number of classifications was assessed both for the training (development) and testing partitions. Since cross-validation was used, the standard deviation of the correct classification accuracy are also measured and reported.

To examine if any of the perturbation techniques did show the Type DM bias, ANOVA is the appropriate test to determine whether significant differences exist in the classification accuracy between the four different treatment groups. If a significant difference is found, then follow-up

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

tests, controlling for multiple comparisons, are appropriate to assess which specific treatment groups (i.e., which perturbation techniques) had significant differences in classification performance. There are a plethora of choices for this test. We opted for one of the most conservative tests, Tukey's HSD.

RESULTS

Tables 1 and 2 illustrate the classification accuracy of the IRIS and LIVER data sets, respectively. Average classification accuracy (over the ten cross-validated trials) is shown (with accuracy standard deviation). The tables show the results by treatment group (Original, SADP, GADP, GADP-W), classification group (three groups for the IRIS Table 1 - IRIS Results (Correct Classification Accuracy) Table 2 - Liver Results (Correct Classification Accuracy) data set, 2 groups for the LIVER set), and are differentiated by training data (constructing the decision tree) and testing data (holdout samples, validating the decision tree).

As an example, in Table 1, the data-mining tool correctly classified an average of 100% of the *Iris setosa* cases, and 96% of both the *Iris virginia* and *Iris veriscolor* cases when using the non-perturbed IRIS data to construct the decision trees (training data) (an overall classification rate of 97.33%). Similarly, and somewhat surprisingly, the test results for the original data show the same average accuracy. When SADP perturbed data were analyzed by QUEST, the classification accuracy decreased to 74.44%, 88.44% and 57.33% respectively for each group (73.41% overall) in the training samples, and 74%, 80% and 50% (68% overall) for the holdout (testing) samples.

When comparing results from the Liver data (Table 2) to the results from the IRIS data (Table 1), we see that the Liver data set was in fact more difficult to classify, as expected.

Using classification accuracy as the dependent measure, ANOVA tests are applied to each individual group of the data set, then for overall total accuracy, for both the training and testing sets (e.g., eight different ANOVA's for IRIS). For the IRIS data, all show significant performance differences (at the <.01 level) among the four treatment groups (F-test values and p-values reported in the Table). Next, Tukey's HSD test was performed to assess which treatment groups differed significantly. These results are also shown in Table 1, but require some additional explanation.

Treatment groups that did not significantly differ from each other appear in the same 'box' in the Table. Likewise, those that significantly differed will appear in different 'boxes'. As an example, consider overall training results for IRIS in Table 1. Three treatment 'groups' differed significantly at the $p<.01$ level. GADP-W and Original differed significantly from SADP, and SADP differed significantly from GADP (and, through transitivity, GADP-W and Original also were significantly different from GADP). Within the first group, GADP-W and Original did NOT significantly differ in their classification accuracy. P-values are shown for all comparisons where significant differences were found.

Similarly, Table 2 reports on the same tests (ANOVA and Tukey's HSD) for the Liver data set. Six ANOVA's are necessary in this case. All training results had significant group differences ($p<.01$ for all situations), but only Group 2 testing results had significant treatment differences. Thus, the Tukey test is only appropriate for the four situations where the ANOVA results showed significant differences.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 1: IRIS Results (Correct Classification Accuracy)

	TRAINING RESULTS			
	Setosa	Virginica	Versicolor	Overall
Original	100.00%	96.00%	96.00%	97.33%
	0.00%	0.94%	0.94%	0.62%
SADP	74.44%	88.44%	57.33%	73.41%
	14.37%	5.22%	25.66%	4.33%
GADP	78.44%	50.22%	58.89%	62.52%
	3.32%	24.62%	18.03%	3.15%
GADP-W	100.00%	100.00%	99.56%	99.85%
	0.00%	0.00%	1.41%	0.47%
ANOVA F-value	34.5	32.8	21.7	456.6
p-value	<.001	<.001	<.001	<.001

	TESTING RESULTS			
	Setosa	Virginica	Versicolor	Overall
	100.00%	96.00%	96.00%	97.33%
	0.00%	8.43%	8.43%	5.62%
	74.00%	80.00%	50.00%	68.00%
	26.75%	16.33%	28.67%	13.63%
	32.00%	76.00%	46.00%	51.33%
	23.48%	20.66%	35.34%	12.98%
	100.00%	100.00%	100.00%	100.00%
	0.00%	0.00%	0.00%	0.00%
	32.5	7.26	15.7	57.5
	<.001	<.001	<.001	<.001

Significant Differences (Tukey HSD test)

GADP-W Original	GADP-W Original	GADP-W Original	GADP-W Original
	SADP		

GADP SADP		GADP SADP	SADP
	GADP		GADP

p<.01 p<.01 p<.01 All p<.01

GADP-W Original	GADP-W Original	GADP-W Original	GADP-W Original

SADP	SADP GADP	SADP GADP	SADP
	GADP		GADP

All p<.01 p<.05 p<.01 All p<.01

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 2: Liver Results (Correct Classification Accuracy)

	TRAINING DATA			TESTING DATA		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Original	84.78%	60.54%	74.59%	72.00%	42.76%	59.71%
	5.07%	8.72%	2.80%	10.85%	14.83%	6.14%
SADP	81.44%	53.03%	69.50%	74.50%	30.34%	55.94%
	5.35%	8.42%	2.28%	12.57%	14.61%	9.07%
GADP	90.00%	36.78%	67.63%	80.00%	20.00%	54.78%
	6.87%	12.23%	2.14%	14.91%	12.78%	5.87%
GADP-W	91.50%	84.06%	88.37%	69.50%	37.24%	55.94%
	1.66%	2.50%	1.14%	14.80%	8.10%	10.85%
ANOVA F-value	8.32	51	185.6	1.1	5.8	0.7
p-value	<.001	<.001	<.001	0.35	0.002	0.57
				n.s.		n.s

Significant Differences

GADP-W GADP	GADP-W	GADP-W	
	Original SADP	Original	
Original SADP		SADP GADP-W	
	GADP		
SADP vs. GADP	p<.01		
Original vs. GADP-W	p<.05	All p<.01	All p<.01

NONE	Original GADP-W SADP	NONE
	GADP	
GADP vs. orig	p<.01	
GADP vs. GADP-\	p<.05	
GADP vs. SADP	n.s.	

The treatment group differences are presented as before, but additional clarification is necessitated due to 'complicated findings.' For Group 1 cases in the training data, significant differences at the p<.01 level were found between GADP and SADP and GADP-W and SADP. Additionally, significant differences at the p<.05 level were detected between Original data and the SADP treatment group and GADP-W (but not GADP).

Evaluating significant differences of Group 2 testing cases requires some explanation as well. Original, GADP-W, and SADP do not significantly differ from each other. GADP and SADP also do not significantly differ, but there is a p<.05 significant difference between GADP and GADP-W and a p<.01 significant difference between GADP and the Original data set.

DISCUSSION

The results provide some evidence that using data perturbation techniques to protect confidential attributes in a database do impact the extraction of non-parametric 'hidden' knowledge-based relationships. However, there are a variety of findings and some unique (and unexpected) results that were observed.

IRIS, the easier data set to 'discover knowledge,' showed fairly predictable results. The original data set was classified by the data mining tool at an almost per-

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

fect 97.33% rate. The data set created by applying the naïve data perturbation approach SADP was classified by QUEST at a relatively poor 68% (testing). Since SADP adds uncorrelated noise to the data to 'protect' it from unauthorized disclosure, this performance is as expected.

The performance of the data mining tool on the data generated by the GADP perturbation process, when GADP did not utilize information about the class (group) variable in the perturbation process, was surprisingly poor, being significantly worse than even SADP. As GADP in general is a more sophisticated approach than SADP and uses more relevant information about the data when creating the protected dataset, this was an unexpected finding. This finding could be an impact of the (arbitrary) manner in which the non-confidential attribute was chosen (lowest correlation with the group variable). This counterintuitive result merits future investigation, and could provide additional insight into the performance of perturbation techniques on data sets with different statistical characteristics. Also, the result might simply be an artifact of the impact of not including the class variable in the perturbation process, and this may not be a realistic view of use in a corporate database.

Equally surprising might be the performance of the data mining tool when classifying GADP-W perturbed data. The accuracy of QUEST (on the IRIS data set) was as good or *better* that of the original data set in every case. Of course, this difference was not statistically significant, so caution must be applied in interpreting these results. In this circumstance, the GADP-W protection method may have acted as a data 'smoothing' technique for the original data set and turned instance 'outliers' into more representative data points.

This 'smoothing' phenomenon did not occur as dramatically for the Liver data. This could be explained by considering the relative 'degree of difficulty' of the classification task. The Liver data set has been shown to be harder to predict than the IRIS data set. Perhaps this 'degree of difficulty' concept impacts the potential 'smoothing' ability of the protection technique. Regardless, data mining proponents have long held that non-parametric procedures such as QUEST are insulated from the effects of outliers. These results may indicate otherwise, and are worthy of future research.

Testing results for the Liver database indicated no significant difference in data mining tool performance when perturbed data or original data was 'mined'. Basically, QUEST did a poor job in classifying Group 2 cases irrespective of the data set used. The results, especially the training data, illustrate that the technique may be maximizing its accuracy by predicting 'Group 1' membership for most of the cases (at the expense of Group 2 cases). The performance of QUEST on the training sets and the overall poor performance at generalizing (test sets) does provide evidence that a Type DM bias may exist. In any event, one can state that the data perturbation techniques have some impact on data mining tool performance.

To summarize, GADP-W perturbed data was surprisingly well classified by QUEST. Likewise, GADP data was surprisingly problematic for the data mining tool. Since SADP is the most naïve data perturbation technique, it is reasonable to expect its performance to be significantly worse than the original data. Given the exploratory intent of the study, the results provide some evidence of a Type DM bias for perturbation schemes for database protection. Given that the two datasets used

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

in this study are on somewhat opposite ends of the 'degree of difficulty' spectrum, future investigation of the impact of data protection schemes on knowledge discovery might look at less 'extreme' datasets. This could lead to a better understanding of the potential Type DM bias, and/or discover the circumstances where the GADP protection technique may actually add additional value to knowledge discovery techniques by reducing the impact of instance outliers.

CONCLUSION

The purpose of this study was to examine whether proposed database perturbation methods useful in protecting organizational databases added a new kind of bias, termed DM or Data Mining bias, to the database. The study looked at the impacts of perturbation techniques protecting two different database scenarios, and the initial results provide some support for the existence of this bias. Other results from this study show evidence of potential 'data smoothing' effects by the 'gold standard' GADP perturbation approach. Ultimately, further research is necessary to explore how characteristics of the data and/or characteristics of the relationship between confidential and non-confidential attributes of the data influence the ability to discover knowledge in our protected database.

As with any study, there are limitations, many of which have been previously discussed. They include the manner in which the database variables were assigned as confidential and non-confidential, since it is possible that the degree of association between the confidential and non-confidential attribute influences both the performance of the data perturbation technique and the knowledge discovery approach.

Also, the conundrum faced by the authors on whether the group (class) label should be utilized in the data perturbation approach could have a big impact of interpretation of the knowledge discovery results (i.e., GADP vs. GADP-W). Thus, we report both results.

Finally, the study used two data sets to represent the spectrum of knowledge discovery difficulty. More exhaustive analysis of different databases is needed to increase external validity. Also, other knowledge discovery techniques (neural networks, other decision tree approaches, etc.) may perform differently, and this warrants further investigation.

In conclusion, the study did find initial evidence that using a perturbation technique to prevent the unauthorized use of confidential attributes may result in an additional 'Type DM' bias. This bias could severely impact the ability of an organization to gain the significant benefits of knowledge management/discovery. All told, the competing needs faced by the DBA to balance data protection and providing a viable data access platform for knowledge discovery activities represent an exciting and relevant line of inquiry as organizations continue to seek differentiable and sustainable competitive advantages.

APPENDIX A—MATHEMATICAL DESCRIPTION OF GADP

Borrowing from the notation used in (Muralidhar, Parsa, and Sarathy, 1999), the GADP can be described as follows:

X represents p number of confidential, numerical attributes

S represents q number of non-confidential attributes

Y represents p number of perturbed attributes

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

The joint covariance matrix of **X** (confidential), **S** (non-confidential), and **Y** (perturbed) can be represented by

$$\Sigma^G = \begin{bmatrix} \Sigma_{XX} & & \\ \Sigma_{XS} & \Sigma_{SS} & \\ \Sigma_{XY} & \Sigma_{SY} & \Sigma_{YY} \end{bmatrix}$$

U = {X, S} with mean vector $\mu_u = [\mu_x, \mu_s]$ and covariance matrix (total set of confidential plus non-confidential attributes combined).

$$\Sigma_{UU} = \begin{bmatrix} \Sigma_{XX} & \\ \Sigma_{XS} & \Sigma_{SS} \end{bmatrix}$$

To simplify, Σ^G can be written as

$$\Sigma^G = \begin{bmatrix} \Sigma_{UU} & \\ \Sigma_{UY} & \Sigma_{YY} \end{bmatrix}$$

where $\Sigma_{UY} = [\Sigma_{XY} \quad \Sigma_{SY}]$

Let $\mathbf{U} = \mathbf{c}_i$ represent the vector of the ith observation set from U. This vector contains observations of both confidential and non-confidential attributes and can be expressed by the following:

$$c_i = [X_{i1,} X_{i2.....} X_{ip,} S_{i1,} S_{i2.....} S_{iq}]$$

To generate the perturbed values for the database, a conditional random vector $(\mathbf{Y}|\mathbf{U} = \mathbf{c}_i)$ is produced. Each perturbed value $\mathbf{Y}|\mathbf{U} = \mathbf{c}_i$ represents the conditional value of Y (perturbed) given a vector of actual observations for **X and S**. The distribution of $\mathbf{Y}|\mathbf{U} = \mathbf{c}_i$ is multivariate normal with expected value and variance (see Graybill, 1976):

$$E(\mathbf{Y}|\mathbf{U} = \mathbf{c}_i) = \mu_x + \Sigma_{YU}(\Sigma_{UU})^{-1}(\mathbf{c}_i - \mu_u)$$
$$\mathbf{Var}(Y|U = c_i) = \Sigma_{YY} - \Sigma_{YU}(\Sigma_{UU})^{-1}\Sigma_{UY}$$

The main addition in the GADP method was the incorporation of the actual values for both the confidential and non-confidential attributes when 'perturbing' the database. Because of this, the relationship between the perturbed confidential values and the non-confidential values are maintained, which removes the Type C bias (while still keeping the other biases from occurring) found in BCADP.

There are certain requirements that must be specified for Σ_{XY} Σ_{SY} and Σ_{YY} so that the four types of bias will be eliminated. Our study follows the guidelines suggested in (Muralidhar, Parsa, and Sarathy, 1999) to implement GADP, including the use of the canonical correlation between the set of confidential and non-confidential attributes (security measure) which impacts the parameter Σ_{XY}.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ira G. Rosen, Laura Rosen and Wayne L. Rosen for their support and contributions to this paper.

REFERENCES

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. North Scituate: Duxbury Press.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. *ASA Proceedings Survey Research Methods*, 370-374.

Lim, T.-S., Low, W.-Y., & Shih, Y.-S. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 40, 203-229.

Low, W.-Y. & Shih Y.-S. (1997). Split Se-

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

lection Methods for Classification Trees. *Statistica Sinica*, 7, 815-840.

Matsumoto, M. & Nishimura, T. (1998). Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1), 3-30.

Merz, C.J. & Murphy, P.M. (1996). *UCI Repository of Machine Learning Databases*. Department of Information and Computer Science, University of California, Irvine, CA,. (http://www.ics.uci.edu/~mlearn/MLRepository.html).

Muralidhar, K., Batra, D., & Kirs. P. (1995). Accessibility, Security and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach. *Management Science*, 41(9), 1549-1564.

Muralidhar, K., Parsa, R. & Sarathy, R. (1999). A General Additive Data Perturbation Method for Database Security. *Management Science*, 45(10), 1399-1415.

Numerical Technologies Random Generator for Excel (NtRand). (2001). *Numerical Technologies Corporation*, 1-21. (http://numtech.com/documents/19981222/index.htm).

Tendick, P. (1991). Optimal Noise Addition for Preserving Confidentiality in Multivariate Data. *Journal of Statistics Planning and Inference*, 27(2), 341-353.

Tendick, P. & Matloff, N (1994). A Modified Random Perturbation Method for Database Security. *ACM Transcations on Database Systems*, 19(1), 47-63.

Weiss, S. & Kulikowski, C. (1991). *Computer Systems that Learn***.** San Mateo: Morgan Kauffman.

Rick L. Wilson is the W. Paul Miller Professor of Business Administration and Professor of Management Science and Information Systems at Oklahoma State University, USA. He received his PhD in Management Information Systems from the University of Nebraska-Lincoln. Dr. Wilson has published numerous papers in the areas of applied artificial intelligence and data mining, applied management science and organizational impacts of emerging technologies.

Peter A. Rosen is a doctoral student at Oklahoma State University, USA, where he is pursuing a degree in Management Science and Information Systems. His past degrees include a BA in Psychology from the University of California, Santa Barbara, and an MBA from San Diego State University. His research interests include the technology acceptance model, meta-analytic and data mining techniques, neural network training algorithms, data security, and group processes. He has previously worked for the OSU MBA Program, Aptex Software, and Unisys.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

View publication stats