

The Impact of Data Perturbation Techniques on Data Mining

Rick L. Wilson, Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK 74078
(405) 744 5084

Peter A. Rosen, Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK 74078
(405) 744 3201

Abstract: Data perturbation is a data security technique that adds 'noise' to databases to allow individual record confidentiality. This technique allows users to ascertain key summary information about the data while preventing a security breach. Four bias types have been proposed which assess the effectiveness of such a technique. However, these biases deal with simple aggregate concepts (averages, etc.) found in the database. In e-commerce applications, organizations are interested in applying data mining approaches to databases to discover additional knowledge about customers.

In our study, we propose a fifth type of bias that may be added by perturbation techniques (Data mining Bias), and empirically test for its existence. Our results find support for this bias, and propose future research avenues that are appropriate for the emerging interdisciplinary field in data security, e-commerce and data mining.

INTRODUCTION

The explosion of data and information captured by organizations about customers, competitors, supply chain partners and internal processes are a well-accepted and recognized phenomenon of the new e-commerce business model. As organizations seek for true personalization in dealing with their customers, the proper utilization of these massive data stores becomes paramount. Simply put, in the e-commerce world, those who best 'mine' data about their customers will come out ahead in the personalization goal of e-commerce applications.

From a data mining or knowledge discovery sense, conventional wisdom indicates that an organization should possess and collect as much information about its customers as possible. However, data privacy and security issues may conflict with this philosophy. This study is interested in a particular form of data privacy/security whereby individual, confidential attribute values are not made available to legitimate users, but aggregate 'relationships' of the database are (hopefully) made accessible. The rationale of these data perturbation techniques is that 'users' do not need to know individual identities of customers (data), only an understanding of aggregate relationships of the customers (data).

Recent studies of data perturbation techniques have proposed increasingly sophisticated approaches to protect confidential data while maintaining the parametric characteristics of the original, non-perturbed (i.e., non-secure) data. For instance, the General Additive Data Perturbation (GADP) method (Muralidhar, Parsa, & Sarathy, 1999) was recently shown to resolve 4 important biases possessed by previously proposed techniques.

However, these biases dealt with simple relationships (means, variances, etc.). Knowledge discovery and data mining techniques so critical for e-commerce applications identify complex, non-parametric relationships from a data set (e.g., decision trees, rules, etc.). Thus it is not known how data perturbation techniques impact these critical 'discovered' relationships. Therefore, this study will empirically explore the possibility of another kind of 'bias', referred to as a Type DM (for data mining) bias.

Specifically, we will empirically examine whether data mining tools' classification accuracy is impaired by perturbation techniques on two 'classic' classification databases, the IRIS and LIVER datasets (Merz & Murphy, 1996). We will build decision trees using the QUEST (Low & Shih, 1997) data-mining algorithm from original data and perturbed data, and test for significant differences in classification accuracy. One might expect that classification accuracy using perturbed data would be lower than classification accuracy on non-perturbed data. However, to our knowledge, there are no prior studies of this phenomenon. Thus, the study will not hypothesize a directional difference a priori. We expect the results of our study to provide a basis for future work at the intersection of the data security, e-commerce, and data mining fields.

REVIEW OF RELEVANT LITERATURE

Data Security through Perturbation Techniques

Databases are ubiquitous and of immense importance to e-commerce applications. The information and knowledge that can be generated from them are absolutely essential; helping organizations to match and customize their products and services to potential customers.

Organizations store large amounts of data, and some (most?) may be considered confidential. Thus, security 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 have legitimate access to the data.

Our interest in data security is focused not on physical and technical access methods, but on statistically based methods that seek to protect confidential data by using data perturbation techniques. Data perturbation involves adding random noise to confidential, numerical attributes, thereby protecting the original data. Even while altering the original data, these methods allow users the ability to access important aggregate statistics (such as means, correlations and covariances, etc.) from the entire database, thus 'protecting' individual records. For instance, in the case of sales data, an employee may not be able to access what a particular individual purchased from a store on a given day, but that employee could determine the total sales volume for the store on the same day.

Muralidhar et al. (1999) examined previously proposed data perturbation methods, and analyzed their effectiveness on bias and security measures. Bias occurs in a data perturbation method when a query generated using perturbed data produces a result significantly different than the same query would using the original data. Four types of biases were identified in (Muralidhar et al., 1999) termed Type A, Type B, Type C, and Type D.

Type A bias occurs when the perturbation of a given attribute causes summary measures of that individual attribute to change due to a change in variance. Type B bias is identified as a bias that occurs when perturbation changes the relationships between confidential attributes. Type C bias occurs when perturbation changes the relationship between confidential and non-confidential attributes. Type D bias deals with the distribution of the data in a database. When the underlying distribution of a given database is not multivariate normal, and/or the added noise term is not multivariate normal, the form of the resulting perturbed database cannot always be determined.

Additionally, so-called security measures may also be important (Muralidhar et al., 1999). Security is measured by the degree to which an unauthorized user can determine the values of confidential attributes in a specific record through the use of the relationships between the non-confidential and confidential attributes. A mathematical way to measure the amount of security provided by a perturbation method is through the use of canonical correlation analysis. Canonical correlation analysis can be used to measure the maximum proportion of the variance that can be explained in any linear combination of confidential attributes, using a linear combination of known (non-confidential) attributes. This study will utilize this measure of security in calculating the perturbed datasets, but remain primarily interested in the bias caused by such methods.

Muralidhar et al. (1999) showed that past methods suffered from one or more of the four aforementioned biases, and thus were inadequate data perturbation techniques. The simplest method, Simple Additive Data Perturbation (SADP) (Kim, 1986), 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, and it can therefore be shown that SADP is inadequate as it suffers from Type A, Type B, and Type C bias.

Other methods proposed include the Correlated-Noise Additive Data Perturbation (CADP) method (Kim, 1986; Muralidhar, Batra, & Kirs, 1995) and the Bias-Corrected Correlated-Noise Additive Data Perturbation (BCADP) method (Kim, 1986; Tendick & Matloff, 1994). These methods are 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 et al., 1995). Unfortunately, this family of perturbation techniques also suffers from all four bias types.

The General Additive Data Perturbation (GADP) method is proposed (Muralidhar et al., 1999) as a further improvement to these past methods and possesses none of the four bias types. The main addition of the GADP method is the incorporation of the actual values for both the confidential and non-confidential attributes when the perturbation process takes place. Because of this, the relationship between the perturbed confidential values and the non-confidential values are maintained, which removes Type C bias found in BCADP.

There are certain requirements that must be specified so that the four types of bias will be eliminated. Our study follows the guidelines suggested in Muralidhar et al. (1999) to implement GADP. More details can be found in Muralidhar et al. (1999).

Proposing the existence of Type 'DM' bias

While the GADP method is constructed to 'secure' confidential attributes appropriately and to eliminate biases to the data, this is a limited view of organizational databases' value-added capability. Knowledge discovery techniques (such as data mining techniques) can identify underlying patterns in a database. Often modeled in decision tree or 'rule' form, these tools help organizations gain deeper insight (knowledge) about their processes, customers and competitors. The biases discussed in Muralidhar et al. (1999) deal only with simple parametric aggregate measures and relationships (means, variances, covariances).

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

Many data mining approaches have been presented and studied (Lim, Low, & Shih, 2000). It is not the intent of our paper to look for the 'best' data mining approach for our experimental circumstances, but to choose a representative approach that will allow us to adequately assess evidence of Type DM bias existence. QUEST (Quick, Unbiased, Efficient Statistical Tree) was chosen based upon its performance in a recent study comparing thirty-three knowledge discovery classification tools (Lim et al., 2000).

METHODOLOGY

Study Procedures – General

Four treatment groups were created. The first treatment group involved analyzing the original data with the data mining tool, while the second and third treatment groups used the SADP and GADP method of data perturbation respectively to perturb the data before the data mining tool was utilized. The final treatment group used the GADP method for data perturbation, but also included the dependent variable as a non-confidential attribute in the perturbation.

In an organizational setting, it is unclear whether the classification group of the individual record would be included in the database and/or known to the user. Thus, we analyzed the GADP method both with and without the categorical dependent variable included in the perturbation. In summary, the four treatment groups are Original, SADP, GADP and GADP-W (with the dependent variable included).

Two frequently used data sets were selected from the UCI Machine Learning Repository to be used as surrogates for organizational e-commerce customer-related data (Merz & Murphy, 1996). The first data set was the IRIS Plant Database, chosen for the almost perfect linear separability of its dependent variable groups. This data set consists of 150 observations, four numerical independent variables (sepal length, sepal width, petal length, petal width) and the categorical dependent variable, class of iris plant. The three levels of the dependent variable are Iris-setosa, Iris-versicolour, and Iris-virginica. 33.3% of the data correspond to each of the three levels of the dependent variable. The three independent variables with the highest correlation to the dependent variable (sepal length, petal length, and petal width) were designated as confidential attributes, and thus perturbed. The remaining attribute, sepal width, was designated as a non-confidential attribute, and was not perturbed.

The second data set utilized in the study was the BUPA Liver Disorders Database (Merz & Murphy, 1996). This data set consists of 345 observations, six numerical independent variables (mcv - mean corpuscular volume, alkphos - alkaline phosphatase, sgpt - alanine aminotransferase, sgot - aspartate aminotransferase, gammagt - gamma-glutamyl transpeptidase and drinks - number of half-pint alcoholic beverages consumed per day), and a dependent variable with two levels, indicating presence or absence of a liver disorder. 145 of the observations were classified as group 1 (42%), and 200 of the observations were classified as group 2 (58%). Five of the six independent variables were considered confidential in this data set (alkphos, gammagt, mcv, sgpt, and sgot), while the final attribute (drinks) was considered non-confidential, and not perturbed. The correlation with the dependent variable was again used as the rule to decide which attributes were to be perturbed. This data set was selected due to the high error rate that other researchers have previously found in classification tests (Lim et al., 2000). Thus, it serves as a complimentary example to the 'easy' IRIS data set.

The data mining tool selected for the study is SPSS's AnswerTree software, using the QUEST method (Kim, 1986) of classification. Ten-fold cross-validation was used to determine a robust measure of classification accuracy for QUEST under each treatment group, and each data set. Ten-fold cross-validation involves splitting a data set into ten equal (or as equal as possible) parts. Nine of the 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. Additionally, the 10 partitions of the data were stratified. The IRIS data set was split into ten parts of 15 observations each, with each of the part containing 5 Iris-setosa, 5 Iris-versicolour, and 5 Iris-virginica observations. The LIVER data set was split in a similar manner, with 34 or 35 observation in each part, with 14 or 15 group 1 classifications and 20 group 2 classifications per part.

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 used for testing during the first run of the data mining tool. Note that a stopping rule was specified for QUEST so that a tree no greater than depth 5 was generated. This was kept constant throughout the study. The next run would use parts 2-10 for training, and part 1 for testing. The process would be repeated until all ten parts of the data were used at one time for the testing set. This is a statistically sound way to determine an accurate measure of classification accuracy (Weiss & Kulikowski, 1991). The correct number of classifications was assessed for both the training and testing sets. If the actual observation, for example, was an Iris-setosa, and the data mining tool classified the flower as Iris-setosa, we would consider that a correct classification. It is now appropriate to formally state the hypotheses of the study:

HO: No Data mining Bias exists; all methods tested produce the same results.

HA: There is a Data mining Bias; at least one method produces results that are different from the other treatment groups.

Perturbation Implementation

SADP was implemented as follows. For each confidential attribute, the mean and standard deviation of the original data was used in the creation of an error term. Using Microsoft Excel's data analysis add-in, random, normally distributed numbers ('errors') were generated with mean 0 and a standard deviation equal to the standard deviation of the original attribute. The values of the original observations for each attribute were added to the corresponding random error term to create the perturbed data set.

The implementation of GADP required numerous tools. First, a multivariate normal distribution random number generator was necessary to implement the procedure. The EXCEL add-in NtRand (Numerical Technologies) was used to generate the necessary perturbed data. NtRand incorporates significant advances in today's psuedo-random number generators. Input parameters were selected to minimize the differences between the original and perturbed covariance matrices of the data set (Matsumoto & Nishimura, 1998; Numerical Technologies). Input into NtRand included the $E(Y|U = ci)$ and $Var(Y|U = ci)$ vectors. It is worth noting that the non-linear EXCEL solver procedure specified in Ashley (1996) was helpful in calculating the canonical correlation between the confidential and non-confidential attributes, a necessary input for the covariance matrix. A VBA application was written which created the GADP and GADP-W datasets using the EXCEL add-ins.

Statistical Assessment

ANOVA is the appropriate test to determine if significant differences exist among treatment groups. If a significant difference is found, then multiple comparison follow-up tests are appropriate to assess which specific treatment groups were significant different from each other in classification performance. While there are a plethora of choices for this test, we opted for one of the most conservative tests, Tukey's HSD. We also examined other, less stringent tests, and found few differences between tests. We report only the Tukey's HSD findings.

RESULTS

(DUE TO SPACE LIMITATIONS, THE RESULTS SECTION IS AVAILABLE UPON REQUEST FROM THE AUTHORS)

DISCUSSION AND CONCLUSION

To some extent, the results were consistent with rejecting the null hypothesis. There appears to be evidence that data perturbation techniques add 'noise' to the data such that underlying patterns of non-parametric knowledge are not as cleanly extracted by data mining techniques as when compared to the original data. However, not all results are consistent, and they differ across data sets.

IRIS, the easier data set to 'discover knowledge', showed fairly predictable results. The original data set was classified at an almost perfect 97.33% rate, while SADP, which just randomly adds noise to the data sets' confidential attributes, saw its classification performance dip to 68% (testing). This was expected. The performance of GADP WITHOUT considering the dependent (group) variable was surprisingly poor, showing results significantly worse than even SADP. Perhaps the manner in which the non-confidential attribute was selected (lowest correlation with the group variable) caused this phenomenon. Future research is necessary to explore the potential impact of the relationship between the canonical correlation coefficient (which is a surrogate for the degree of relationship between confidential and non-confidential attributes) and Type DM bias.

Equally surprising might be the performance of GADP-W. The accuracy of QUEST (on the IRIS data set) was at least as good or better than that of the original data set in every case. (Of course, this difference was not statistically significant, which would support the null hypothesis.) In this circumstance, perhaps the GADP-W method acted as a data 'smoothing' technique for the original data set and turned data 'outliers' into more representative data points. This phenomenon did not occur as dramatically for the Liver data, perhaps a function of the relative 'degree of difficulty' in the classification task. Either way, further research into this phenomenon is warranted. Traditional data mining research has long held that the non-parametric procedures deal with outliers implicitly and are not subject to their effects like parametric procedures; perhaps GADP-W (or GADP in general) has a heretofore-unknown capability as an outlier reduction technique for data mining.

Testing results for the Liver database indicated no significant difference between the data perturbation techniques and the original data. Basically, QUEST did a poor job in classifying Group 2 cases irrespective of which data set it used. The results, especially for 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). While some of the testing results support the null hypothesis, 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. Alternatively, one can state that the data perturbation techniques have some impact on data mining tool performance.

To summarize, GADP-W data was surprisingly well classified by QUEST. Likewise, GADP was a surprisingly poor performer. 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 do support two lines of further inquiry: 1) There is evidence of a Type DM bias and 2) There is evidence that an approach such as GADP-W may be useful in increasing data mining accuracy when a data set possesses 'outliers'.

This study did find initial evidence that a Type DM bias does exist, and perhaps other phenomenon which data perturbation techniques place on knowledge discovery processes. As an introductory study, perhaps more questions were posed than were answered. Nonetheless, the need for both strong knowledge discovery tools and data privacy and security in the quest for continued e-commerce success makes this interdisciplinary research area worth pursuing in much more detail.

ACKNOWLEDGEMENTS

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

REFERENCES

Ashley, D. W. (1996). A Canonical Correlation Procedure for Spreadsheets. *Proceedings, 27th Annual Meeting of the Decision Sciences Institute*, 1102-1109.

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 Selection 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*. Retrieved July 14, 2001 from University of California, Irvine, Department of Information and Computer Science site: <http://www.ics.uci.edu/~mlearn/MLRepository.html>

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

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.

Numerical Technologies Random Generator for Excel (NtRand). Retrieved July 17, 2001, from Numerical Technologies Corporation site: <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 Transactions on Database Systems*, 19(1), 47-63.

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