Global Co-Operation in the New Millennium
The 9th European Conference on Information Systems
Bled, Slovenia, June 27-29, 2001

DATA MINING PROTOTYPE FOR DETECTING E-COMMERCE FRAUD [RESEARCH IN PROGRESS]

Monkol Lek, Benjamin Anandarajah, Narciso Cerpa, Rodger Jamieson

The University of New South Wales, Sydney, NSW 2052, Australia
Tel.: 61 2 9385 4414, Fax: 61 2 9662 6041
mlek@au1.ibm.com , ben.anandarajah@accenture.com, n.cerpa@unsw.edu.au, r.jamieson@unsw.edu.au

ABSTRACT

The advent of electronic commerce (e-commerce) has marked a significant change in the way business now approach the implementation of their sales and marketing strategies. Electronic commerce growth has been accompanied by an increase in fraudulent practices. Researchers have proposed rules-based auditing systems for electronic commerce transactions, but they highly depend on the auditor's knowledge of e-commerce fraud (Wong et al., 2000). While fraud patterns may occur, the management, control, and application of these patterns is difficult due to the increasing number of online transactions currently handled by e-commerce systems. In this paper, we present research in progress of the prototype of an extension to e-commerce auditing systems that use data mining techniques to generate rules from fraud patterns. Subsequently, the system applies these rules to e-commerce databases with the aim of isolating those transactions that have a high chance of being fraudulent.

1. INTRODUCTION

Whereas it was once acceptable for companies to sell their products to very defined and localised markets within certain logical timeframes, the advent of online shopping has completely redefined the way companies now market themselves in order to establish a market presence (Nath et al.,1998). However, the introduction of this dynamic medium of conducting business has brought with it it's own complex set of problems. Although many businesses are well placed to be able to capture the emerging markets that electronic commerce can open up, factors such as widespread concerns about fraud and Internet security have greatly hindered online business prospects (Smith, 1999). It must be noted that these concerns are shared by both consumers as well as corporate organisations, who stand to lose sizable amounts from fraudulent activities (Sweeney, 1999).

Although many varied approaches to fraud prevention are undertaken by organisations (Baker, 1999), the lack of any coordinated and uniform approach to the solution of these problems has meant that the various deficiencies that exist across networked systems can be systematically exploited. The use of systems which are able to build user profiles and then use this knowledge to extract patterns of fraud, will be better able to adapt as patterns of attack inevitably evolve. Practical application of the technologies of data mining and detection algorithms, will allow companies to identify and understand patterns of fraudulent behaviour, and then take effective action to combat these incidences (Lach, 1999).

This paper presents research in progress on the development of a data mining prototype for detecting fraud patterns and irregularities in electronic commerce transactions. These transactions are vulnerable to fraud since they are done remotely and often it is quite difficult to verify the legitimacy of customers. This fraud detection software prototype provides auditors and IS management with the appropriate tools for detecting, managing, and controlling fraud patterns, as well as applying the rules derived from these patterns to commercial e-commerce databases. The software prototype provides a graphical user interface for the merchants' fraud investigators to determine the data to be considered in the rule generation process. The following sections provide an outline of the theoretical foundation for fraud pattern detection and continuous auditing then discuss the design and implementation of the e-commerce fraud detection prototype software and its usefulness to the business, auditors and IS management.

2. RESEARCH OBJECTIVES

The aim of this research is to understand fraud detection expertise and to develop a software prototype that will detect fraudulent e-commerce transactions using data mining techniques. Our goal is to provide a tool which is simple and intuitive to use, and which allows IS management and auditors to classify data sets without exposing them to the low-level implementation details of the software.

The vast majority of currently available security systems focus only on very specific applications of their software, and do not try to implement a more portable and extensible approach to the prevention of online fraud. As a result, this research endeavors to develop a fraud detection model and implementation, such that a common framework for the detection of electronic commerce fraud could be used across a variety of industries to detect incidences of fraudulent behaviour.

One of the major problems facing organisations is the inordinate amount of overhead associated with initially configuring a fraud prevention package for a particular system, as well as the ongoing resources necessary to constantly fine-tune the software. Since the fraud detection prototype is implemented in Java, it's 'write once, run anywhere' capability provides flexibility and the capacity to support a variety of database and operating system standards, using the many widely available device drivers. In addition, by providing an interface to constantly learn from training sets of data, the task of updating it's detection capabilities will be simplified. An intuitive Windows-based graphical user interface will allow a system administrator to set the exact parameters for fraud detection, which may vary greatly across organisations and industries.

One major hurdle for developers of intrusion detection systems has been the inability of systems to react rapidly enough to detect when changes to patterns of attack have occurred. In many cases, by the time an attack has been detected, it is far too late for an administrator to be able to track down the offender and determine exactly what damage has occurred (Lunt, 1993). As a result, any fraud detection software must be able to work within very defined time constraints, for it to be of commercial use.

The use of artificial intelligence (AI) techniques may automate the detection processes that system administrators would intuitively go through, when searching for patterns of behaviour that may be later classed as fraudulent (Lach, 1999). The research investigated many algorithms and possible implementations of fraud pattern detection software in order to develop a software prototype capable of efficiently and accurately detecting instances of electronic commerce fraud from a company's transactional data, and is also able to evolve to changing patterns of attack. Some of the algorithms and implementations considered included decision tree implementations such as ID3 (Mitchell, 1997), C4.5 (Quinlan, 1993) and Ripple Down Rules (Compton et al., 1991), as well as the closely related field of intrusion detection systems.

To summarize, the aims of the research are first to enhance IS practice by the construction of a fraud detection prototype to provide a tool which is able to interface to a range of commercial e-commerce databases. The second aim is to contribute to IS theory by building a model of e-commerce fraud detection.

3. THEORETICAL FOUNDATION

Organisations world-wide who are able to integrate data mining processes into their every day operations stand to benefit greatly from the insights gained into applications as diverse as customer relationship management, the optimisation of marketing and advertising techniques, as well as the detection of fraudulent behaviour in electronic commerce environments (Lach, 1999).

While still a relatively young area of research, data mining, and the discovery of various techniques of implementation, have undergone considerable advances in the last few years, as both commercial and research organisations aim to take advantage of the substantial insights into business management that datamining can provide (Groth, 2000). One of the most active, and possibly pervasive fields of research in the past few years, has been the discovery of algorithms for the use of analysing sets of data and extrapolating meaningful information from them. The field relies on the combination of statistical and artificial intelligence knowledge, in order to process data in an efficient and accurate manner (Holsheimer et al.,1994). Purely statistical methods were often not robust enough to handle the fairly dynamic nature of data which is generated by industries such as financial services (Elder et al., 1996). Data sets usually contained a large number of incorrect or incomplete information, which made the task of deducing non-trivial knowledge extremely difficult. As a result, a number of approaches had to be developed to handle the diverse types, and quality of data sets from which knowledge was to be gained.

The implementation of data mining system to extract non-trivial knowledge is usually carried out in one of three ways, using data association rule algorithms, sequential analysis, or classification rule-learning approaches (Fayyad et al., 1996). Although a great deal of research is currently being carried out into the applications of the association rule algorithms and sequential analysis for fraud detection, our research focusses on the application of classification rule-learning algorithms for this purpose. We feel that this approach captures the intent and functionality of such a tool, which can be used by organisations to easily process vast amounts of transactional data in order to search for fraudulent transactions.

4. RESEARCH METHODOLOGY

The research methodology consists of the following phases: literature review; investigation of AI algorithms; construction of early prototype using ID3; testing of the prototype; development of the datamining prototype using C4.5; testing of the datamining prototype; discussions with fraud investigators and the review of fraud cases in order to develop e-commerce training sets; field testing of the prototype using an e-commerce organisation against its commercial databases, and development of a research model for e-commerce fraud. Since our e-commerce fraud detection model should be able to work as an independent module and also to fit within the e-SCARF framework (Wong et al., 2000), see Figure 1, both these requirements were considered within the design and implementation.

Following completion of the literature review, the authors investigated different artificial intelligence techniques currently used for intrusion detection and electronic fraud detection with the aim understanding their pros and cons. Since machine learning (classification algorithms) have proven to be one of the most successful artificial intelligence techniques (Groth, 2000, Lach, 1999), this research implements a machine learning algorithm. A prototype to demonstrate how a classification algorithm (i.e. ID3 (Mitchell, 1997)) can be used in conjunction with a database was built. The ID3 algorithm is fairly closely related to our preferred algorithm, C4.5, some of the main differences being the ways that the calculation of gain and tree pruning algorithms are implemented. Both of these functions are substantially less sophisticated, and as a result, the implementations are much less involved than for C4.5. The major difference however is that ID3 can only process discrete-valued attributes, while C4.5 has the ability to deal with both discrete and continuous attribute values, which is essential for data associated with e-commerce transactions.

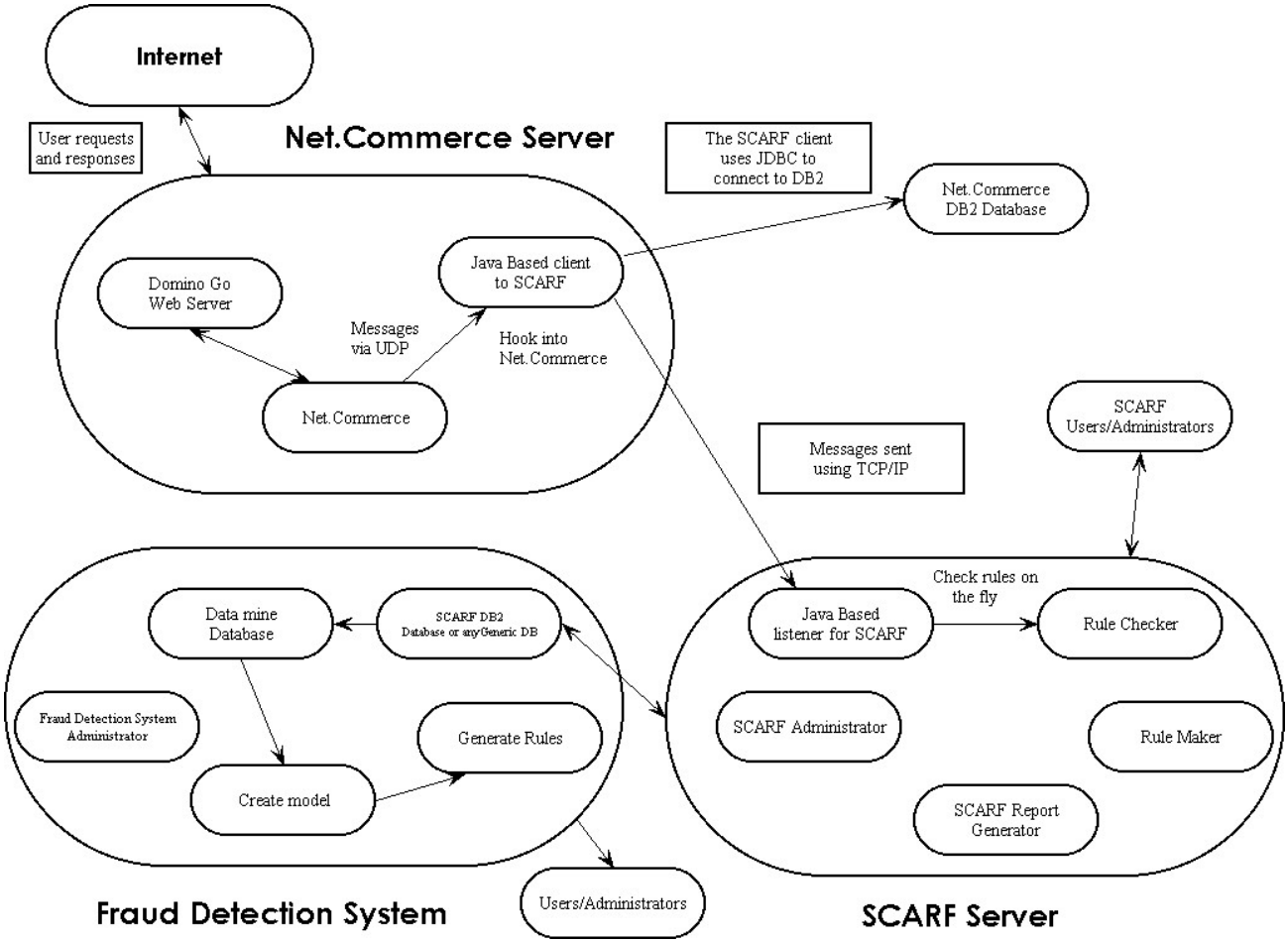

Figure 1: *Interaction between datamining prototype and e-SCARF (System Control Audit Review File) for e-commerce fraud detection*

The main motivation for building the early prototype was primarily to observe ID3 at work, to analyse difficulties in this implementation, and then use this information when designing and implementing the final software version using C4.5, a far more ambitious and complicated algorithm. Finally, it provides a very convenient basis for the benchmarking and testing of the early prototype on standard data sets. From the comparisons undertaken on the implementations, further improvements and modifications will then implemented.

5. PROGRESS TO DATE

The design procedure for this prototype was rapid prototyping with no formal method of requirements analysis nor specification of system functionalities being carried out. In hindsight, this approach caused a number of problems - the most substantial of these being the production of code which was overly large, and thus difficult to manage. As a result, a strict object oriented design methodology is planned to be followed for the implementation of the C4.5, final version of our software, with the primary aim being the production of highly readable, portable, and most importantly, extensible code.

The early prototype was implemented in Java over a span of 3 weeks. The Java implementation placed most of the emphasis on the coding of the decision tree rather than the coding of the GUI, which resulted in the production of somewhat non-elegant code for the graphical user interface. The prototype functionality consisted of:

- Aesthetically pleasing GUI (yet simple and primitive)
- Connections to either locally or remotely MySQL database using JDBC technology

- Displays for database information and other available catalogs
- Displays for table information such as table names, attribute names, attribute sizes, etc
- Simple SQL query executions and display of the results in a formatted JTable
- Construction of a decision tree based model on data from a particular table in a database
- Model testing using another table with the same attributes in the database
- Simple naive pruning of the tree

The Prototype was evaluated using data sets from the University of California, Irvine's Machine Learning data set repository. It was tested against a more advanced decision tree, namely C4.5. The C4.5 code used, was the Quinlan implementation in C, which is provided in his book (Quinlan, 1993). The results are outlined below:

Data Set	**Training size**	**Test size**	**Quinlan C4.5 (Pre)**	**Quinlan C4.5 (Post)**	**Prototype ID3**
Monk1	124	432	76.6	75.7	75
Monk2	169	432	65.3	65	48.6
Monk3	122	432	92.6	97.2	87.3

Table 1: *Comparison of early prototype with other algorithms*

Target Concepts: The Monk's problems are three artificial tasks designed to test learning algorithms. (Thrun et al., 1991)

It should be noted here that the prototype did not implement decision tree pruning. Although the results did not show this, the trees created by the prototype were much larger than the pruned trees of C4.5. Since the Prototype did not perform pruning, it thus performed relatively poorly when compared to C4.5. Both C4.5 and the Prototype performed quite poorly on the Monk2 data set because the concept to be learnt was quite complex in nature, and was better dealt with by using a neural network approach for this case.

Some of the difficulties with the prototype implementation were as follows:

- How to efficiently and conveniently store each sample so that it can be retrieved and inserted with ease. The samples were simply stored in a Hashtable of their own, with it's attribute values being the keys for retrieval and insertion. The Samples were then stored in a vector.
- How to organise the data that forms the decision tree, such that the pruning which was carried out later was achieved in a simple and efficient manner.
- How to implement a pruning technique that was algorithmically efficient, and capable of actually restructuring the tree, rather than merely pruning off subtrees.

For the final version of our software, the C4.5 classification algorithm, which was proposed by (Quinlan, 1993), was chosen. The C4.5 algorithm generates a classification-decision tree for a given data set by the recursive partitioning of data. The C4.5 algorithm is an extension of the earlier algorithm ID3, which was used to implement our early prototype. C4.5 was used for the later prototype as it is able to deal with training sets that have records with unknown attribute values, by evaluating the gain, or the gain ratio, for an attribute by only considering only the records where that attribute is defined. As a result records that have unknown attribute values are classified by estimating the probability of the various possible results. The ability to handle missing values is of critical importance for the domain of electronic commerce data, where most the values are input by humans, and as such, are likely to contain errors and omissions. Secondly, unlike ID3, C4.5 is able to process continuous-valued attributes, a fairly important requirement for systems whose main sources of data would originate from commercial applications. The inability to process continuous-valued samples would necessitate having to manually discretise any samples to be classified.

6. CONCLUSIONS

Our model and its initial implementation provide the support required by practitioners according to the literature, since it is very intuitive to use, it does not require a great deal of customisation, and it can learn from new fraud instances. The prototype fits very well within the e-SCARF framework (Wong et al., 2000), and is also able to work as an independent tool. The prototype has been demonstrated to IS audit management who have been impresed by its potential. Trials will now take place with an ecommerce organisation where detailed investigations will be undertaken with fraud investigators to provide the training set cases and then to run the prototype against the e-commerce databases. Results from these trails will also be discussed with the State and Federal police computer crime units who have agreed to cooperate in this research and who are very interested in this fraud detection technique.

REFERENCES

Baker, C. R. (1999) An Analysis of fraud on the Internet, *Internet Research-Electronic Networking Applications & Policy*, 9(5), 348-359.

Compton, P., Edwards, G., Kang, B., Malor, R., Menzies, T., and P. Preston (1991) Ripple down rules: possibilities and limitations, *Proceedings of the 6th Knowledge Acquisiting for Knowledge Based Systems Workshop*, Banff, pp 2-5.

Elder IV, J. and D. Pregibon (1996) *A Statistical Perspective on Knowledge Discovery in Databases*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, pp. 83-115, AAAI/MIT Press.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and R. Uthurusamy (1996) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

Groth, R. (2000) *Data Mining- Building Competitive Advantage*, Prentice Hall.

Holsheimer, M., and A. P. J. M. Siebes (1994) Data Mining, the search for knowledge in databases, *Report CS-R9406*, CWI, Amsterdam, The Netherlands, pp. 8-19, 41-49.

Lach, J. (1999) Data Mining Digs In, *American Demographics*, July, pp38-40, 42-45.

Lunt, T. L. (1993) A Survey of Intrusion Detection Techniques, *IPIP-TC11 Computers and Security*, 12(4), pp 405-418.

Mitchell, T. M. (1997) *Machine Learning*. Singapore: McGraw-Hill.

Nath, R., Akmanligil, M., Hjelm, K., Sakaguchi, T., and M. Schultz (1998) Electronic Commerce and the Internet: Issues, Problems, and Perspectives, *International Journal of Information Management*, 18(2), pp 91-101.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Smith, R. (1999) Fraud : What Response?, *Australian CPA*, November, pp. 39.

Sweeney, P. (1999) Cyber-Crime's Looming Threat, *Banking Strategies*, July/August, pp 54-56,58-59.

Thrun, S. B., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., De Jong, K., Dzeroski, S., Fahlman, S. E., Fisher, D., Hamann, R., Kaufman, K., Keller, S., Konomenko, I., Kreuziger, J., Michalski, R. S., Mitchell, T., Pachowicz, P., Reich, Y., Vafaie, H., Van de Welde, W., Wenzel, W., Wnek, J., and J. Zhang (1991) The Monk's problems: A performance comparison of different learning algorithms, *Technical Report CMU-CS-19-197*, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.

Wong, K., Ng, B., Cerpa, N., and R. Jamieson (2000) An Online Audit Review System for Electronic Commerce, *Proceedings of the Thirteen Bled Electronic Commerce Conference*, Slovenia, pp 19-21.