Journal of Forensic & Investigative Accounting
Vol. 1, Issue 1

Data Mining Techniques for Auditing Attest Function and Fraud Detection

John Wang
James G.S. Yang*

The need for data mining in the auditing field is growing rapidly. As the online systems and the hi-technology devices make accounting transactions more complicated and easier to manipulate, the use of data mining in the auditing profession has been increasing in recent years. Since auditing involves evaluation of massive data in the attest function, data mining allows this process to be done in an easier manner. Auditors use computer aided audit software (CAATs) to make the process more accurate and reliable.

There are three basic approaches to data mining: mathematical-based methods, distance-based methods, and logic-based methods [1] The first approach, mathematical-based methods, uses neural networks, which are networks of nodes modeled after a neuron or neural circuit that mimics the human brain. These neural networks are used in the auditing profession in many different ways, such as risk assessment, finding errors and fraud, determining the going concern of a company, evaluating financial distress, and making bankruptcy predictions. The next approach to data mining is distance-based methods which uses clustering to put large sets of data into groups and classifications based on attributes. This method is commonly used in marketing but is also useful in auditing. The third approach to data mining is the logic-based approach which uses decision trees to organize data. In the field of auditing, the logic-based method is most

* The authors are, respectively, Professor of Operations Research and Professor of Accounting at Montclair State University.

commonly used. Specifically, it can be efficiently applied to the analysis of bankruptcy, bank failure, and credit risk. Data mining approaches can make the auditing function easier by organizing and analyzing data in a more efficient and effective way.

Continuous vs. Periodical Auditing

Technology improvements have changed the way auditing is being performed. The traditional financial auditing is performed periodically but ironically, financial data are continuously flowing through electronic circuit. Therefore, the traditional auditing function is being threatened by the use of information technology system. To solve this problem, more and more auditing firms have started to use continuous auditing. It is "a methodology that enables independent auditors to provide written assurance on a subject matter using a series of auditors' reports issued simultaneously within a short period of time after the occurrence of events underlying the subject matter."[2] Since so many transactions are being recorded electronically without the use of paper documentation, continuous auditing allows for "real-time assurances from an independent third party that the information is secure, accurate, and reliable."[3] Mr. Shire, the CEO of PriceWaterhouseCoopers, said that "the Internet, stakeholders' demands for real-time financial information, new corporate value drivers, global stock trading, 24-hour business news, and security needs for electronically transmitted information are fundamentally changing the way we do business. The demand for information that is on time and accurate is forcing the accounting profession to rethink how auditors audit their companies. Investors and other users of financial reports are beginning to demand more timely and forward-looking information, which will mean that continuous auditing will replace the traditional year-end report."[4] This indicates that nowadays, an auditor's role

is no longer a periodic activity, but a continuous on-line performance. Data mining is one of the tools that make continuous auditing a possibility. It is particularly useful in performing the audit attest function on an on-going basis. As continuous auditing starts to replace the traditional auditing, the data mining techniques will undoubtedly be used by auditors more and more.

Forensic Audit as a New Function

In addition to continuous auditing, there is a new category of auditing that is currently in the making. Investors expect auditors to detect all frauds, but the auditing firms are not the insurers or last resort of the capital markets. The auditors' liabilities are not comparable to the auditing fees the investors are willing to pay. Therefore, there is an "expectation gap" between investors and auditors. On November 8, 2006 the heads of all big four accounting firms plus the next two biggest firms issued a policy paper entitled "Serving the Global Capital Market and the Global Economy,"[5] in which it suggests the institution of a system that requires companies to be subject to "forensic audits" specifically designed to root out fraud and corporate wrongdoing every three to five years. This is a new category of audit in addition to the regular financial audit. It will definitely cost more for the companies, but it may greatly reduce auditors' liabilities. As a result, the investors' interest may also be better protected. Data mining is purposely designed for fraud detection. It makes continuous auditing possible. It can also make forensic auditing a reality.

Application of Data Mining in the Auditing Profession

One major area of auditing is the making of going-concern predictions about a company. The auditors are required by auditing standards to assess the status of a company and make a prediction as to whether the company is able to continue operating as a going concern. Determining the going concern status of a company is a very difficult task, so auditors have been trying to come up with statistical methods to help make it easier. In the article "Going Concern Prediction Using Data Mining Techniques,"[6] 165 going concern companies and 165 non-going concern companies were used in a study to assess the effectiveness of data mining techniques in determining the going concern of a company. Decision trees, neural networks, and regression were used to test the sample. The results found that the usefulness of data mining to predict a company's going concern was very high. The decision tree model had an accuracy rate of 95%, the regression model 94%, and the neural network model 91%.[7] All three models were able to predict which companies were going concerns. Data mining is changing the way auditing is being performed by adding information technology into audit services and providing the opportunity to improve audit effectiveness.

FRAUD DETECTION

Employee theft costs $40 billion a year with an average of $1,350 per employee, of which 75% of them are undetected.[8] The total losses of shoplifting amount to $13 billion a year with 800,000 incidents each day costing $142 loss per incident. However, only 5,000 incidents were caught.[9] Identity theft is also becoming very serious. In the past five years there were 27.5 million victims of identity theft. In 2006 alone there were 9 million victims. The losses were $48 billion to businesses and financial institutions and $9 billion to consumers per year.[10] These losses are ever increasing today.

The above losses prove a point that detecting fraud is a constant challenge for any business. Implementation of data mining techniques has been shown to be cost effective in many business applications. It is particularly useful in the field of auditing, such as fraud detection, forensic accounting and security evaluation. "Randall Wilson, director of fraud at RGL in St. Louis, agreed that the growth in computer forensics has been nothing short of incredible, especially in the area of employee misappropriation. He has picked up countless cases of collusion between employees and outside vendors, complete with fraudulent invoices. Clearly, there has been an increase in the opportunities for fraud and, consequently, increased opportunities for catching fraud. Wilson explained that what has happened in the business world has triggered a rise in fraudulent activities. As a result, his company is doing more data mining, simulation, fraud detection and prevention."[11] This indicates that data mining has played an important role in today's business operations.

Applications of Data Mining in Fraud Detection

There are two applications in data mining that can be used to detect fraud: Outlier Analysis and Benford's Law Analysis. In Outlier Analysis the data which are very different from the rest of the data (outliers) are identified. The outliers can be the result of errors or something else like fraud. This analysis identifies these deviations that are not the norm and have a higher risk of being fraudulent. Benford's Analysis is a technique that allows the auditor to quickly assess the data in ways that will detect

potential variances. Benford's Law was named after Dr. Frank Benford, who discovered that, within a large enough universe of numbers that were naturally compiled, the first digits of the numbers would occur in a logarithmic pattern. This analysis concludes that, if numbers do not follow the Benford pattern, then something abnormal must have happened with the data, which could lead to the detection of a fraud.

In direct application of effective data mining techniques, here are some examples of results of fraud detection:

1. Discovery of a packaging supplier being paid over $4 million and not supplying any products to the company,
2. Discovery that a vendor was issuing fraudulent invoices on a regular and sequential basis which indicated that the vendor had only one customer,
3. Discovery of payments to family members by government officials, and
4. Discovery of a senior executive who issues invoices to a fraudulent company with his own home address.

Data Mining in the Department of Defense

Here is an example of data mining techniques in action. The Defense Contract Audit Agency (DCAA) is responsible for performing all contract audits for the Department of Defense (DoD), in addition to providing accounting and financial advisory services regarding contracts and subcontracts to all of the DoD. In recent years, DCAA's IT group has developed data mining software tools to assist their auditors in analyzing

contractor data. This is not an off the shelf application, but rather an application developed in-house. This data mining software was developed to help improve the efficiency and accuracy of their audit of large government contractors that use the Deltek System 1 or GCS Premier accounting systems. These applications are MS Access based and can handle the largest of corporate files. This tool is a menu driven application which imports a standard set of tables and creates standard reports in pivotal table format.

In addition, the Defense Finance and Accounting Service (DFAS) utilized data mining analysis a few years back. The DFAS provides responsive and professional finance and accounting services for the DoD. Since it is responsible for disbursing nearly all of the DoD funds, they implemented data mining techniques to minimize fraud against DoD assets. They selected SPSS Inc.'s Clementine data mining software to implement the financial service.[12] In the end, DFAS's data mining analysis selected payments for further investigation of fraud.

Data Mining in the Health Insurance Industry

Insurance fraud has become very widespread. It costs $80 billion a year, translating into $875 additional cost to each person. Health insurance fraud alone costs $30 billion a year. Medicare insurance fraud is worse. It costs $179 billion a year.[13] Auto insurance fraud losses are 12.3 billion a year, adding $200 premium to each driver.[14] Property insurance fraud costs $30 billion a year.[15] And the losses are still rising. These losses indicate an urgent need for fraud detecting techniques. Another example of data mining in a real world situation is as follows:

Example: In line with Megaputer,[16] having to deal with millions of insured customers and thousands of providers and medical services, healthcare insurance companies have to routinely address the task of verifying the authenticity and legitimacy of all transactions that are processed. In fact, fraudulent transactions might originate from all involved parties. For instance:

a. Fraudulent "providers" can be sharing lists of valid patient IDs and trying to bill the insurance company for the services that were never rendered in reality.
b. Fraudulent "patients" might try to bill insurance companies for a large number of ghost procedures.

As fraud schemes get more sophisticated and the volume of transactions grows, it becomes increasingly more difficult to discern fraudulent from legitimate transactions. Auditors have to utilize advanced data mining tools capable of processing large volumes of data and detecting unusual events that deviate from the normal operation patterns. Fraud schemes are rapidly changing and analysts need to be able to discern new fraud patterns without an explicit prior knowledge of these patterns.

Using a combination of *PolyAnalyst* data mining techniques, valuable results can be obtained in the search for traces of the two possible healthcare fraud mechanisms listed above. A large health insurance company provided the following dataset containing about 15,000 records representing individual procedure transactions. This is a subset of

the patient records of a health insurance company for the year 2004. It contains only those patients who had more than 160 procedures during the year. The important fields of data are listed below: *PatientID, PatientName, ProviderID, ProviderName, ServiceDate, ProcedureCode, NetPayments*. It can launch Summary Statistics exploration engine first in order to gain an overall picture of the data. A positive Link Chart selecting *PatientName* as antecedent attribute and *ProcedurCode* as consequent attribute can be created. The minimum correlation number can be selected. It then locates those patients who have the largest number of links to different procedures and evaluates some of these links in more details. One of the most frequent manifestations of fraud of the second type might be the presence of the same service obtained by a patient on the same date from the same or different providers. All performed procedures were quite expensive. Now it becomes really suspicious about the relationship between some patients and their providers. As a result, we might want to carry out a much more in-depth analysis to determine whether we have indeed identified some fraudulent behavior. Surprisingly enough, we discovered the first example of suspicious behavior not for those patients who had the largest number of procedures performed during the year 2004, but for one of the patients who had about the smallest number of procedures compared to other patients in the selected dataset. This is a very typical situation in fraud detection projects. While it is often hard to tell for certainty whether we have discovered a fraud or just some peculiar behavior, this analysis can certainly help auditors identify the situations where a red flag should be raised on certain patients or provider, and thus suggests that the corresponding transactions should be scrutinized further.

Data Mining in the Department of Homeland Security.

After 9/11 under the U.S.A. Patriot Act the banking institutions are required to report not only transactions of suspected money laundering and illegal drug trafficking activities, but also potential terrorist operations to the U.S. Treasury Department's FinCen. Banks now adopt "defense filing" – 281,373 in 2002 to 689,414 in 2006. Only 5,000 out of 1,100,000 reports since 9/11 are terrorists related.[17] Nevertheless, data mining techniques can narrow down the search for terrorists.

Data Mining in Efficient Deployment of Investigative Resources

In yet another example, consider the problem of fraud investigation, perhaps in the area of insurance claims. Suppose some 10,000 cases have been investigated and of those just 5% (or 500) were found to be fraudulent. This is a typical scenario for numerous organizations. With modeling we wish to improve the deployment of our resources so that the 95% of the cases that were not fraudulent need not all be investigated, yet the 5% still needs to be identified. Each case of actual fraud also has a dollar value associated with it, representing the magnitude of risk associated with the case.

The advantage of *decision tree*, a more specific tool of data mining, is that the resulting tree (and particularly if we traverse each path through the tree to obtain a set of rules) can be easily understood and explained. It provides decision makers with an opportunity to understand the changes being suggested.

Williams[18] found that, if our investigators actually only investigated 25% of the cases that they are currently investigating and then they would recover 64% of the cases that were found to be fraudulent and that 72% of the dollars were recovered. Thus, the other

75% of the investigative resources could be better deployed, perhaps in higher risk populations where the returns are greater. Note that the *strike rate* has increased from 26% in the original dataset to 67% at this optimal point. With half of the resources currently deployed on investigations (i.e., a caseload of 50%). A better result can conceivably be obtained by using the data mining model that can recover almost 90% of the fraud and recover more than 90% of the dollar amount.

Data Mining in Sarbanes-Oxley Act

In another area of application, the debacles of WorldCom, Enron, Adelphia, Xerox and HealthSouth have revealed some of the largest accounting cover-ups in history. As a consequence, the Sarbanes-Oxley Act (SOX) was enacted in 2002. The Act was aiming at higher corporate governance standard, and increasing transparency, accuracy and integrity of corporate financial reporting.

The mission of SOX is noble. Unfortunately, its compliance is costly. In a survey of corporate chief financial officers, it was found that "the average cost of complying was $1.7 million for companies with market value ranging from $75 million to $699 million. Companies with a market value greater than $700 million reported average compliance cost of $5.4 million in 2005."[19] According to a study by one industry group, "companies on average spent $3.8 million each in fiscal 2005 to comply."[20] In another survey of corporate executives, it was estimated for all companies "that companies would spend $6 billion on compliance with the rules in 2006, down only slightly from $6.1 billion in 2005."[21] In fact, the actual cost of compliance was $2.92 million in 2006.[22] The SOX compliance cost is indeed tremendous. Nevertheless, there is one way to reduce it.

The auditors can use many different tools and technologies to analyze financial data. "Analysis products that can assist with Sarbanes-Oxley compliance consist of querying, data mining, and financial statement examination tools. Each of these tools is designed to facilitate analysis of organizational data to identify risks that may not be apparent on the surface and can be used to validate that controls are effective."[23] This indicates that the data mining techniques can be employed as tools in complying with SOX. Further, since the techniques are so cost-effective, they can greatly reduce the SOX compliance cost.

The above examples demonstrate that data mining techniques have wide areas of applications in business operations, such as banks, insurance companies, telecom, airline companies, credit companies, etc. Even the police, the FBI and the Homeland Security all employ data mining techniques to detect crimes and terrorists' activities. These operations have to be able to discern fraudulent transactions from the main stream of legitimate business transactions. Inability to catch fraud can become an extremely painful, damaging and costly problem to a business. Therefore, the need for efficient fraud detection solutions resides high on the "to do" list of every company.

DATA MINING FOR IMPROPER PAYMENTS

Improper payments are a widespread and significant problem in governments and private businesses. These payments include inadvertent errors, such as duplicate payments and miscalculations, payments for unsupported or inadequately supported claims, payments for services not rendered, payments to ineligible beneficiaries, and payments resulting from outright fraud and abuse by program participants and/or

employees. For example, in the federal government, improper payments occur in a variety of programs and activities, including those related to contractors and contract management, health care programs, such as Medicare and Medicaid, financial assistance benefits, such as Food Stamps and housing subsidies, and tax refunds. The causes for improper payments are many, ranging from fraud and abuse, poor program design, inadequate internal controls and simple mistakes and errors.

In the private sector, improper payments most often present an internal problem that threatens profitability, whereas in the public sector they can translate into serving fewer recipients or represent wasteful spending or a higher relative tax burden that prompts questions and criticism from Congress, the media, and the taxpayers. For federal programs with legislative or regulatory eligibility criteria, improper payments indicate that agencies are spending more than necessary to meet program goals. Conversely, for programs with fixed funds, any waste of federal funds translates into serving fewer recipients or accomplishing less programmatically than could be expected.

The Office of Management and Budget (OMB) has estimated that at least $35 billion is improperly spent each year. That represents approximately 10 percent of the non-defense discretionary budget authority requested in the FY 2005 budget. The Deputy Director of OMB said recently that just whittling away at improper or erroneous payments could save the federal government $100 billion over the next 10 years.[24]

Data mining analyzes the data for relationships that have not previously been discovered. As a tool in managing improper payments, applying data mining to a data warehouse allows an organization to efficiently query the system to identify questionable activities, such as multiple payments for an individual invoice or to an individual recipient on a certain date. This technique allows personnel, who are not computer specialists but may have useful program or financial expertise, to directly access data, target queries, and analyze results. Queries can also be made through data mining software, which includes prepared queries that can be used in the system on a regular basis.

The challenges associated with using data mining to address the problem of improper payments include establishing a data set of known fraudulent payments, a target population of non-fraud, and a method by which to leverage the known fraud cases in the training of detection models. As is typical in fraud detection, the set of known cases was very small relative to the number of non-fraud examples. Thus, researchers have to devise methods to reduce the false alarms without drastically compromising the sensitivity of the models.

The first step is to obtain the data needed to perform the analysis. For the most part, the actual transactions are used. However, for some other transactions, source documents may have to be used to recreate those transactions. The results are a data set of fraudulent payment candidates that will be used to develop models predicting similar transactions. The challenge for the data mining effort is to predict suspicious payments

using a very small set of known fraudulent payments relative to a larger population of non-fraudulent payments.

The next step is to transform the data. In an effort to identify the vendor payment fraud, the experts have hypothesized dozens of potentially useful transformations of known information that might serve as useful indicators of fraud. Examples of data transformations made in this step include setting flags that identify:

1. Payments addressed to P.O. Box or Suite.
2. Invoices from the same vendor paid to multiple addresses.
3. Invoices from multiple vendors paid to the same address.
4. Invoices from the same vendor which are not sequential based on date submitted.
5. Vendor's addresses matches employees addresses.
6. Highest paid vendors on a comparative basis.
7. Changes in aggregate amounts paid to vendors over time.
8. Payments made under various approval limits.
9. Payments of employee salaries and bonuses not in agreement with master file data or to terminated employees.

Although a single fraud/not-fraud binary label for the output variable can be used, multiple fraudulent payment types can be identified to comprise the different styles of payments in the known fraud data.

The third step is to analyze the relationships and patterns in the data by application software. The different levels of analysis that are available in data mining are artificial neural networks, genetic algorithms, decision trees, nearest neighborhood method, rule induction, and data visualization. In general, the relationships sought are classes, clusters, associations, and sequential patterns. These relationships allow the data to be mined according to predetermined groups, logical relationships, and associative relationships. This allows the data to be mined according to certain criteria, i.e. when improper payments are likely to occur or what categories of vendors are more likely to receive improper payments. This also allows for the prevention of improper payments by mining the data to anticipate patterns and trends.

GENERAL AUDITING SOFTWARE VERSUSS DATA MINING SOFTWARE

The General Auditing Software

The most common software that auditors use is generalized audit software (GAS). It provides for data manipulations, risk assessment, high-risk transaction and unusual items, continuous monitoring, fraud detection, key performance indicators tracking, and standardized audit program generation. However, although audit features, such as sorting, querying, aging, and stratifying are built into GAS packages, auditors are still required to observe, evaluate, and analyze the results. Examples of GAS software include Audit Command Language (ACL), Interactive Data Extraction and Analysis (IDEA), DB2 Intelligent Miner for Data, DBMiner, Microsoft Data Analyzer, SAS Enterprise Miner, SAS Analytic Intelligence, and SPSS. The most popular GAS package that is purchased by auditors is Audit Command Language (ACL) because it is

convenient, flexible, and reliable. ACL is commonly used for data-access, analysis and reporting. The interactive capability of ACL allows auditors to test, investigate, and analyze results in a short period of time. Auditors can easily download their client's data by connecting their laptops to the client's system for further processing. This allows the auditor to view the client's files, steps, and results at any time. Similar to other GAS software, ACL is not able to deal with complex data. ACL does have an Open Data Base Connectivity (ODBC) to reduce this problem; however some files are still too intricate. As a result, auditors face control and security problems.

Although GAS is widely used by auditors today, data mining can provide these users with more extensive conclusions. Data mining software offers auditors automated capabilities to discover useful information. The software has the ability to handle complex problems that are limited by the human brain. Data mining is scalable and can handle an unlimited amount of data in the data warehouse or any size problem. Data mining can uncover interesting information hidden in the accounting transactions that when performing normal work, auditors may not come across. It can be used even when the auditors do not know what they are looking for.

Using data mining also has some drawbacks. Data mining software requires substantial technical skills. The auditor should be able to understand the differences among various types of data mining algorithms so as to choose the right one to use. They should possess the ability to use the software and interpret the results. Although data mining is useful to handle complex problems, sometimes the complexity of the outcome

is too difficult for auditors to understand. Also, since data mining is done automatically, it is difficult to determine how the system came up with the results. This is a major problem for auditors. Another problem that auditors find when using data mining software is the lack of interface among different data mining algorithm methods. The software tends to focus on a single method and utilize only a few techniques that cannot integrate with other software. Finally, although data mining is becoming cheaper, it is still expensive compared to other software. Besides paying for the software itself, users must also include the cost of preparing the data, analyzing the results and training auditors to use the software.

Types of Data Mining Software

There are several data mining software packages that auditors can use. The software can be classified according to their level of sophistication that range from low-end to high-end data mining tools. The more sophisticated data mining tools can handle more complex tasks by using multiple methods and algorithms including wizards and editors for data preparation and can incorporate scalability and automation. The low-end data mining tools are not difficult to use and provide the capability to query, summarize, classify, and categorize data. However, the software is not sophisticated enough to recognize patterns.

The high-end data mining software include CART, WizSoft, Clementine, Enterprise Minder, and Oracle Darwin. These tools are used in complex cases with an enterprise-

scale database management system, such as Oracle or DB2. Oracle Darwin is mostly used for activity-based costing, cost-benefit analysis, and credit analysis, while Enterprise Minder and Clementine are primarily used by marketing companies for trend analysis, customer retention, and product/market analysis. Despite being used by marketing companies, Clementine is used by auditors for fraud detection and credit scoring. Although Clementine is a complex data mining tool, it has a visual programming interface that simplifies the data mining process.

Classification and Regression Trees (CART) is used by auditors to assess the financial risk of a business entity. The auditors use CART to find hidden patterns in data to develop decision trees that can be used to predict the entity's financial risk. Based on these results, the auditor can predict the likelihood that a business will fail, as well as the overall business risk of trading partners, corporate affiliates, investment partners, and takeover targets.

WizSoft is software based on mathematical algorithms and is used for both data mining and data auditing. It features six products: WizWhy, WizRule, WizSame, WizDoc for Office, WizDoc for Web, and WizCount for Reconciliation. WizWhy is a data mining tool that is used for fraud detection. The software utilizes patterns of previous cases of fraud to detect new fraud incidents. WizRule is an audit and cleansing application software that reveals the rules in the data and automatically indicates auditing rules that are being broken. WizSame reveals records that are duplicated such as duplicate payments, two customer names that differ by one letter or two addresses that

are synonymous. "WizCount bank and account reconciliation reveals all the matching transactions, thus leaving out the non-reconciled records. WizCount makes use of several sophisticated mathematical algorithms that quickly cover the enormous number of one-to-one, one-to-many and many-to-many matching possibilities, and reveal the right ones." [25]

Microsoft Excel is an example of low-end data mining software that is used with database systems to build assessments. It is used for a variety of audit applications, including tests of online transactions, sampling, internal control evaluation, and specialized fraud procedures. Special software add-ins, such as risk and sensitivity analyzers, can be used to make accounting management easier. Also, PivotTables can be created in Excel to summarize large amounts of data.

Using any of these data mining packages can assist auditors with intricate transactions in large volumes. As transactions are made, recorded and stored electronically, all of the tools are capable of capturing, analyzing, presenting and reporting the data. Manipulating complicated data through data mining gives auditors the opportunity to analyze information that is beyond their human capabilities. As a result, the auditing market presents tremendous opportunity for an explosive growth of data mining integration.

GUIDELINES FOR USING DATA MINING

The auditors who use data mining as a tool for creating competitive business intelligence should follow several important guidelines to successfully use the software.

First, the auditor should always start with a goal that will provide a solution to the business problem. Secondly, it is important that the data is in the proper format for data mining. Preparing the data is a time-consuming task but a very necessary activity because many times the data received from the data warehouse or data mart are in the wrong format for data mining. Next, a learning sample should be created to use directly in building the model and a testing sample should be developed to evaluate the model for data mining. Fourth, it is important for the auditors to have some basic knowledge of the model-building process. Model building "is a computer-intensive activity that requires both an understanding of the business problem and the data mining methodology for building the model."[26] Novice auditors should begin by using low-end tools that provide easy-to-use assistance such as add on tools in spreadsheet programs, e.g., Excel, and tools that uses highly intuitive graphical use of interfaces. Techniques that are easy to interpret should be used, such as clustering, regression models, and decision trees. Finally, after building the data mining model, auditors should evaluate and validate it to assess the likelihood that it will work by using the testing sample. The effectiveness of different techniques should be compared to find the one that produces the most accurate results.

CONCLUSION

Data mining is growing in the auditing field each day. As technology continues to progress, the use of data mining will continue to benefit the accounting profession. This paper points out that data mining is useful in all areas of accounting, such as auditing, fraud detection, and improper payments. Whether it is through neural networks, genetic algorithms, decision trees, nearest neighborhood method, rule induction, or data visualization, data mining organizes the data in such a way that it makes the accounting

task easier. This paper further identifies many types of data mining software that auditors can use. As data mining continues to grow, there will be more applications in the accounting profession.

References

1. New York State Society of Certified Public Accountants, "Continuous Auditing, XBR and Data Mining," 2005. www.nysscpa.org/committees/emergingtech/auditing2004.ppt .
2. Zhao, N., D. C. Yen, and I. Chang, "Auditing in the E-Commerce Era," *Information Management & Computer Security,* 2004, 12 (5), p. 389.
3. Ibid.
4. Ibid.
5. The Wall Street Journal, "Auditing Firms Urges New Ways to Detect Frauds," November 8, 2006, page C3.
6. Koh, H. C., "Going Concern Prediction Using Data Mining Techniques," *Managerial Auditing Journal,* 2004, 19 (3), p. 462.
7. Ibid.
8. Paulsell, Mary, "The Problem of Employee Theft," University of Missouri, Small Business Development Center, October 1, 2002.
9. Case, John, CPP, "Stopping the Shoplifters," Business Security Publications, 2004.
10. U.S.A. Today, March 2, 2004.
11. Kahan, S., "Bring 'Em Back Intact!" *Accounting Technology,* 2005, p. 16.
12. Data Mining – Clementine Software (Software), 2005. www.the-data-mine.com/bin/view/Software/ClementineSoftware
13. Wikipedia Foundation, "False Insurance Claims," a free encyclopedia, 2006.
14. Wagner, Abby, "Insurance Fraud is a Crime – And It's Costing Consumers Money," InsWeb, 2007.
15. Insurance Information Institute, Inc., "Insurance Fraud," 2006.
16. Megaputer, (2006), Megaputer Intelligence. http://www.megaputer.com/
17. Wall Street Journal, July 7, 2005, page C1.
18. Williams, G. (2006). Predicting fraud: Underrepresented classes. http://www.togaware.com/datamining/survivor/Predicting_Fraud.html
19. The Wall Street Journal, "Small Firms' Sarbanes Suffering?" April 6, 2006, page C1.
20. The Wall Street Journal, "Business Wins Its Battle to Ease a Costly Sarbanes-Oxley Rule," November 10, 2006, page A1.
21. The New York Times, "Top Regulator Says Sarbanes-Oxley Act Are Too Costly and Inefficient," December 1, 2005, page C4.
22. The Wall Street Journal, "Costs Fall Again for Firms to Comply With Sarbanes," May 16, 2007, page C7.
23. Lanza, R. B., "Making Sense of Sarbanes-Oxley Tools," *The Internal Auditor,* p. 48.
24. United States General Accounting Office, "Financial Management Strategies to Manage Improper Payments at HUD, Education, and Other Federal Agencies, GAO 03-1671," 2002.
25. WizSoft – Data and Text Mining, 2005. www.wizsoft.com. www.auditsoftware.net/community/expo2004/Present/23.ppt

www.dcaa.mil
26. Calderon, T. G., J. J. Cheh, and I. Kim, "How Large Corporations Use Data Mining to Create Value," *Management Accounting Quarterly*, 2005, p. 1-13.