Data Mining for Fraud Detection: Toward an Improvement on Internal Control Systems?

Mieke Jans, Nadine Lybaert, Koen Vanhoof

Abstract

Fraud is a million dollar business and it's increasing every year. The numbers are shocking, all the more because over one third of all frauds are detected by 'chance' means. The second best detection method is internal control. As a result, it would be advisable to search for improvement of internal control systems. Taking into consideration the promising success stories of companies selling data mining software, along with the positive results of research in this area, we evaluate the use of data mining techniques for the purpose of fraud detection. Are we talking about real success stories, or salesmanship? For answering this, first a theoretical background is given about fraud, internal control, data mining and supervised versus unsupervised learning. Starting from this background, it is interesting to investigate the use of data mining techniques for detection of asset misappropriation, starting from unsupervised data. In this study, procurement fraud stands as an example of asset misappropriation. Data are provided by an international service-sector company. After mapping out the purchasing process, 'hot spots' are identified, resulting in a series of known frauds and unknown frauds as object of the study.

1 Introduction

Fraud is a million dollar business and it is increasing every year. "*45% of companies worldwide have fallen victim to economic crime in 2004 and 2005. The average damage to the companies from tangible frauds (i.e. asset misappropriation, false pretences, and counterfeiting) was US$ 1.7 million.*" according to the 'Global economic crime survey 2005' of PriceWaterhouseCoopers. Journal headlines and news topics indicate the same trend of increasing fraudulent behavior. Given these numbers, it is remarkable that 34% of these frauds is detected by chance. This gives us a glimpse of the state detection models are in.

Fraud is detected in many ways, or at least one tries to detect it in many ways. Traditionally, a company relies most on its internal control activities and the internal auditor to prevent and detect fraud. If this isn't sufficient, external audit (as far as this isn't legally enforced yet), risk management systems, a whistle-blowing hotline, an investigations department, new technologies or other measures are installed, corresponding to the need the company experiences. The technological improvement that's partly responsible for the increasing trend in fraud, is also part of the solution. Prevention and detection technologies are implemented, tested, customized and commercialized. Software companies sell 'The solution to all your business problems, including fraud'. Also governments use one liners like 'the newest weapon to defeat fraud'. The term 'data mining' is sold as an expensive, all-problems-solving word. If your business doesn't use data mining, you're not in the game.

If this was really the case, then why is there still fraud? Because using data mining or machine learning technologies implies a lot of conditions. First, the term data mining is used many times in an improper manner. Most importantly, data mining is different from the traditional data analysis techniques. Second, the most promising results in fraud detection by means of data mining are attained with supervised learning. Having labeled data is however not a realistic view on most of company's problems. Third, the success stories (that are certainly present!) all address external fraud, give or take a few. The fraud most companies want to combat however is internal fraud.

Internal control systems seem to be an appropriate mean to combat internal fraud, since it is number two (after accidental detection) in detecting fraud. But, it still is number two. This has to be improved, and if we believe part of all the success stories of data mining and fraud detection, we have a candidate for the desired improvement: data mining techniques. This study evaluates the added value of data mining techniques to internal control systems, which are currently merely reporting tools.

2 Theoretical Foundations of the Study

In this section four topics will be covered. An introduction about fraud will be given. What is fraud, how can it be classified and is it worth talking about? Internal control will be highlighted after fraud. Are internal control systems sufficient as a fraud detecting mechanism? In a third part, the topics machine learning and data mining are covered. What makes an analysis fall under these terms and what is the difference with reporting? After clarifying these questions, we turn in the last part to two classes of machine learning, supervised and unsupervised learning.

2.1 Fraud

2.1.1 What is Fraud?

There are many definitions for fraud, depending on the point of view considering. According to *The American Heritage Dictionary, Second College Edition*, fraud is defined as *'a deception deliberately practiced in order to secure unfair of unlawful gain'*. Davia et al. (2000) paraphrase this in a number of items that must be identified, when articulating a case of fraud:

- a victim
- details of the deceptive act thought to be fraudulent
- the victim's loss
- a perpetrator (i.e., a suspect)
- evidence that the perpetrator acted with intent
- evidence that the perpetrator profited by the act(s)

In a nutshell, "fraud always involves one or more persons who, with intent, act secretly to deprive another of something of value, for their own enrichment" (Davia et al., 2000). Wells (2005) stresses *deception* as the linchpin to fraud. To exclude kinds of fraud we don't wish to examine, the delineation of fraud to 'occupational fraud and abuse', as referred to by the Association of Certified Fraud Examiners, is of interest. Occupational fraud and abuse may be defined as: "*The use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets.*" (ACFE, 2006) This definition encompasses a wide variety of conduct by executives, employees, managers, and principals of organizations. Violations can range from asset misappropriation, fraudulent statements and corruption over pilferage and petty theft, false overtime and using company property for personal benefit to payroll and sick time abuses. (Wells, 2005)

2.1.2 Classifying Fraud

The delineation of fraud to 'occupational fraud and abuse' is a good start to study the desired scope of fraud. Yet still, a further classification is convenient. There are numerous ways of classifying occupational fraud. The classification most used is the one where two types of fraud are distinguished: financial statement balance fraud and asset-theft fraud. The main difference between the former and the latter is that there is no theft of assets involved in the former. (Davia et al., 2000) Bologna and Lindquist (1995) classify fraud on many ways, amongst them fraud for versus against the company, internal versus external fraud, management versus non-management fraud and transaction versus statement fraud. Some of them overlap the above mentioned classification into financial statement balance fraud and asset-theft fraud. For example, asset-theft fraud will be fraud against the company and transaction fraud, without being classified as internal, external, management or non-management fraud. Various combinations can be made in this manner.

2.1.3 Some Numbers...

Two elaborate surveys, one in the United States (ACFE, 2006)[1] and one worldwide (PWC, 2005)[2], yield the following information:

45% of companies worldwide have fallen victim to economic crime in the years 2004 and 2005. No industry seems to be safe and bigger companies seem to be more vulnerable to fraud than smaller ones. Small businesses however suffer disproportionate fraud losses. The average financial damage to companies subjected to the PWC survey, was US$ 1.7 million per company. Participants of the ACFE study estimate a loss of 5% of a company's annual revenues to fraud. Applied to the estimated 2006 United States Gross Domestic Product, this would translate to approximately US$ 652 billion in fraud losses for the United States only.

Regarding to the types of fraud, asset misappropriation was number one in both studies. In the PWC survey, this was followed by financial misrepresentation and corruption, false pretences, insider trading, counterfeiting and money laundering. The ACFE report handles a different classification, where asset misappropriation takes 91% of the reported cases for its account, corruption 31% and fraudulent statements 11%.[3]

About the way fraud is detected, both studies stress the importance of tips and chance in detecting fraud. According to the ACFE report, an anonymous fraud hotline anticipates a lot of fraud damage. In the cases reviewed, organizations that had such hotlines, suffered a median loss of US$ 100.000, whereas organizations without hotlines had a median loss of US$ 200.000. At the PWC study, no less than 34% of the fraud cases was detected by means of tip-offs and other 'chance' means. Internal audit and internal control systems can have a measurable impact on detecting fraud

[1] 1.134 cases of occupational fraud, reported by a Certified Fraud Examiner between January 2004 and January 2006, are subject of this report

[2] 3.634 companies around the world are subjected to the Global Economic Crime Survey 2005

[3] The sum of the percentages exceeds 100% because several cases involved schemes that fell into more than one category

after chance related means. The more control measures a company puts in place, the more incidents of fraud it will uncovered.

2.2 Internal Control as Fraud Detection Mechanism

Talking to employees of international companies about fraud detection, many of them answer "We have a very good internal control system. Fraud is not possible here.". This is a common range of thought. What is meant with internal control, will depend on who is asked. Generally speaking, internal control implies a system of well designed processes and procedures for the purpose of fraud prevention and deterring.

Are internal control systems sufficient as a fraud detection mechanism? Apparently not, since over one third of the fraud cases in the surveys are discovered by chance. Internal controls can be split into two groups: active and passive internal control systems. Active internal controls are signatures, passwords, segregation of duties etc. As Davia et al. (2000) put, these can be compared with fences. They may appear insurmountable at first sight, but like all fences, they have their weakness to be defeated by clever fraud perpetrators. And like a fence, once evaded, there is little or no continuing value in preventing or deterring fraud. (Davia et al., 2000) Passive internal controls operate at a different level. Instead of *preventing* fraud, like active controls attempt to, the emphasis here is on *deterring*. Passive internal control systems induce a state of mind in the would-be perpetrator that strongly motivates him "not to go there". Examples of passive control systems are surprise audits, customized controls and audit trails. Passive control systems, when turned active if a company feels the need to do so (they suspect fraud), mainly make use of reporting tools, like providing different numbers and statistics for manual analysis.

Neither active nor passive control systems are best. They complement each other and should both be prevalent.

2.3 Machine Learning and Data Mining

The current information age is overwhelmed by data. More and more information is stored in databases and turning these data into knowledge creates a demand for new, powerful tools. Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. These interpretations and insights are the sought knowledge. So although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts. (Michalski et al., 1998)

To overcome the above limitations, a data analysis system has to be equipped with a substantial amount of background knowledge, and be able to perform reasoning tasks involving that knowledge and the data provided. (Michalski et al., 1998) In effort to meet this goal, researchers have turned to ideas from the machine learning field. This is a natural source of ideas, since the machine learning task can be described as turning background knowledge and examples (input) into knowledge (output). By doing so, the emergence of a new research area was set and frequently called data mining and knowledge discovery. (Michalski et al., 1998)

According to Witten and Frank (2000), data mining can be defined as

> *"... the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities."*

This definition validates Michalski et al. (1998)'s explanation. If data mining results in discovering meaningful patterns, data turns into information. Information or -in this case- patterns that are novel, valid and potentially useful are not merely information, but knowledge. One speaks of discovering knowledge, before hidden in the huge amount of data, but now revealed. This brings us to the term 'Knowledge Discovery', which is usually called in the same breath as 'Data Mining'.

Where we have seen that data mining is a way of discovering knowledge in substantial databases, traditional data analysis techniques merely summarize data and provide important insights. It is important to keep this difference in mind when one speaks of data mining. Governments, Non Governmental Organizations (NGO's), companies and most importantly software suppliers often show of with the term data mining, while they actually implement a traditional data analysis technique. Therefore, this study looks beyond the salesmanship and tries to find out if data mining is really such a success story as is declared everywhere. The amazing possibilities of data mining viewed apart are clear, but is it a realistic assumption that it is also the appropriate solution to real world fraud detection? That's the question.

2.4 Supervised versus Unsupervised Learning

After clarifying the terms machine learning and data mining, it is worth looking at literature using these techniques for the purpose of fraud detection. The machine learning and artificial intelligence solutions that are explored, may be classified into two categories: 'supervised' and 'unsupervised' learning. In supervised learning, samples of both fraudulent and non-fraudulent records are used. This means that all the records available are labeled as 'fraudulent' or 'non-fraudulent'. After building a model using these training data, new cases can be classified as fraudulent or legal. Of course, one needs to be confident about the true classes of the training data, as this is the foundation of the model. Another practical issue is the availability of such information. Furthermore, this method is only able to detect frauds of a type which has previously occurred. In contrast, unsupervised methods don't make use of labeled records. These methods seek for accounts, customers, suppliers, etc. that behave 'unusual' in order to output suspicion scores, rules or visual anomalies, depending on the method. (Bolton and Hand, 2002)

Whether supervised or unsupervised methods are used, note that the output gives us only an indication of fraud likelihood. No stand alone statistical analysis can assure that a particular object is a fraudulent one. It can only indicate that this object is more likely to be fraudulent than other objects.

In what follows we give an overview of the explored data mining techniques for fraud detection, divided into supervised and unsupervised techniques. This overview takes only data mining tools, and no reporting tools or traditional data analysis techniques, into account. Furthermore, it is restricted to mentioning the technique used, without elaborating on the practical decisions the authors made. For a more detailed overview, we refer to Phua et al. (2005) and Bolton and Hand (2002). First the supervised methods used in the literature will be listed, then the unsupervised.

2.4.1 Supervised Methods of Fraud Detection

The use of supervised methods of data mining for fraud detection is investigated in several studies. An intensively explored method are neural networks. The studies of Barson, Field, Davey, McAskie, and Frank (Barson et al.), Fanning and Cogger (1998) and Green and Choi (1997) all use neural network technology for detecting respectively fraud in mobile phone networks (Barson et al.) and financial statement fraud. Lin et al. (2003) apply a fuzzy neural net, also in the domain of fraudulent financial reporting. Both Brause et al. (1999) and Estévez et al. (2006) use a combination of neural nets and rules. The latter use fuzzy rules, where the former use traditional association rules. Also He et al. (1997) apply neural networks in the supervised component of their study. (For the unsupervised part they use Kohonen's Self-Organising Maps) A Bayesian learning neural network is implemented for credit card fraud detection by Maes et al. (2002) (aside to an artificial neural network), for telecommunications fraud by Ezawa and Norton (1996) and for auto claim fraud detection by Viaene et al. (2005).

In the same field as Viaene et al. (2005), insurance fraud, Major and Riedinger (2002) presented a tool for the detection of medical insurance fraud. They proposed a hybrid knowledge/statistical-based system, where expert knowledge is integrated with statistical power. Another example of combining different techniques, can be found in Fawcett and Provost (1997). A series of data mining techniques for the purpose of detecting cellular clone fraud is used. Specifically, a rule-learning program to uncover indicators of fraudulent behavior from a large database of customer transactions is implemented. From the generated fraud rules, a selection has been made to apply in the form of monitors. This set of monitors profiles legitimate customer behavior and indicate anomalies. The outputs of the monitors, together with labels on an account's previous daily behavior, are used as training data for a simple Linear Threshold Unit (LTU). The LTU learns to combine evidence to generate high-confidence alarms. The method described above is an example of a supervised hybrid as supervised learning techniques are combined to improve results. In another work of Fawcett and Provost (1999), Activity Monitoring is introduced as a separate problem class within data mining with a unique framework.

Another framework presented, for the detection of healthcare fraud, is a process-mining framework by Yang and Hwang (2006). The framework is based on the concept of *clinical pathways* where structure patterns are discovered and further analyzed.

The fuzzy expert systems are also experienced with in a couple of studies. So are there Pathak et al. (2003), Bordoni et al. (2001) and Deshmukh and Tallur (1997).

Stolfo et al. and Lee et al. delivered some interesting work on intrusion detection. They provided a framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. Next to this, the results of the JAM project are discussed. JAM stands for Java Agents for Meta-Learning. JAM provides an integrated meta-learning system for fraud detection that combines the collective knowledge acquired by individual local agents.

Cahill et al. (2000) design a fraud signature, based on data of fraudulent calls, to detect telecommunications fraud. For scoring a call for fraud its probability under the account signature is compared to its probability under a fraud signature. The fraud signature is updated sequentially, enabling event-driven fraud detection.

Rule-learning and decision tree analysis is also applied by different researchers, e.g. Shao et al. (2002), Fan (2004), Bonchi et al. (1999) and Rosset et al. (1999).

Link analysis comprehends a different approach. It relates known fraudsters to other individuals, using record linkage and social network methods. (Wasserman and Faust, 1998) Cortes et al. (2002) find the solution to fraud detection in this field. The transactional data in the area of telecommunications fraud is represented by a graph where the nodes represent the transactors and the edges represent the interactions between pairs of transactors. Since nodes and edges appear and disappear from the graph through time, the considered graph is dynamic. Cortes et al. (2002) consider the subgraphs centered on all nodes to define communities of interest (COI). This method is inspired by the fact that fraudsters seldom work in isolation from each other.

2.4.2 Unsupervised Methods of Fraud Detection

The use of unsupervised learning for fraud detection is not explored as intensively as the use of supervised learning. Bolton and Hand are monitoring behavior over time by means of Peer Group Analysis. Peer Group Analysis detects individual objects that begin to behave in a way different from objects to which they had previously been similar. Another tool Bolton and Hand develop for behavioral fraud detection is Break Point Analysis. Unlike Peer Group Analysis, Break Point Analysis operates on the account level. A break point is an observation where anomalous behavior for a particular account is detected. Both the tools are applied on spending behavior in credit card accounts.

Also Murad and Pinkas (1999) focus on behavioral changes for the purpose of fraud detection and present three-level-profiling. As the Break Point Analysis from Bolton and Hand, the three-level-profiling method operates at the account level and it points any significant deviation from an account's normal behavior as a potential fraud. In order to do this, 'normal' profiles are created (on three levels), based on data without fraudulent records. In this respect, we better use the term semi-supervised instead of unsupervised. To test the method, the three-level-profiling is applied in the area of telecommunication fraud. In the same field, also Burge and Shawe-Taylor (2001) use behavior profiling for the purpose of fraud detection. However, using a recurrent neural network for prototyping calling behavior, unsupervised learning is applied (in contrast to Murad and Pinkas (1999)'s semi-supervised learning). Two time spans are considered at constructing the profiles, leading to a current behavior profile (CBP) and a behavior profile history (BPH) of each account. In a next step the Hellinger distance is used to compare the two probability distributions and to give a suspicion score on the calls.

A brief paper of Cox et al. (1997) combines human pattern recognition skills with automated data algorithms. In their work, information is presented visually by domain-specific interfaces. The idea is that the human visual system is dynamic and can easily adapt to ever-changing techniques used by fraudsters. On the other hand have machines the advantage of far greater computational capacity, suited for routine repetitive tasks.

With Bolton and Hand, Murad and Pinkas (1999), Burge and Shawe-Taylor (2001) and Cox et al. (1997), the most important studies concerning unsupervised learning in fraud detection are quoted. Although this list may not be exhaustive, it is clear that research in unsupervised learning with respect to fraud detection is due for catching up.

3 Research Questions

The theoretical background reveals interesting research opportunities. Summarizing the above, companies worldwide have a disastrous problem, costing them a lot of money. The problem calls fraud, more specifically occupational fraud. Through its occupation, one can misuse an organization's assets for personal enrichment, and apparently people do so. The most occurring fraud seems to be asset misappropriation. After uncovering fraud by chance or tip-offs, internal control can have a measurable impact on detecting fraud. Active and passive control systems complement each other well. Yet, internal control comes second in detecting fraud, after accidental detection. Hence, there is room for improvement.

Knowing that internal control systems are currently especially products of reporting tools, data mining could offer a solution in improving and updating certain existing internal controls. If we believe some software selling companies, it is even the best cure against fraud. However, the term data mining is often used for nothing more than standard reporting. If in a following step, literature is reviewed about the trials of using real data mining techniques for fraud detection, it appears researchers have already succeeded in this intention in a promising way. Most of those promising studies involve supervised data. This is however not a realistic representation of the situation most companies are in. Moreover, success stories are in consumer fraud, not in occupational fraud.

Taking all this background information together, it would be interesting to do some research about how to improve existing internal control systems, that currently rely on reporting. In this light, we believe an investigation on the use of data mining for the purpose of occupational fraud detection, starting from a real world assumption, namely unsupervised data, forces itself on. Since occupational fraud encompasses still a very wide range of frauds, it is best to focus on asset misappropriation, since this is threat number one. The research questions that are put forward are:

> "Is data mining, started from unsupervised data, an appropriate solution for detecting asset misappropriation?" If yes,
> "Which data mining techniques are effective in detecting asset misappropriation, starting from unsupervised data?"

These research questions are formulated to the end of solving the internal control topic: *Can data mining mean an improvement for existing internal control systems?* In the following section, a research design for these questions is formulated.

4 Research Design

Davia et al. (2000) compare the art of fraud detection, with the art of fishing.

..., expert fishermen never simply go fishing for fish. Rather, they first decide what type of fish they have a taste for. Next, they decide the how, with what equipment, and where they will expertly search for that type of fish and that type alone.

Following this advice of first deciding what sort of fraud you are looking for, the asset misappropriation fraud has to be narrowed down. In this study, we will search for procurement fraud, as an example of asset misappropriation.

Data will be provided by and of an international service-sector company, willing to cooperate. The company, Epsilon named in this study, is of considerable magnitude. It employs 55.000 people around the world, of which 40.000 in the Benelux. As compared to a manufacturing-sector or merchandizing-sector company, a service-sector company will purchase for smaller amounts of money. Yet, Epsilon purchases for around 1.26 billion euros each year, a considerable amount.

As a start, the purchasing process within Epsilon is audited. This is done by reviewing internal procedures, users guides and audit reports, by interviewing persons in charge of relevant departments and by following executives in their job. Once the purchasing process is mapped, 'hot spots' were identified. Out of these hot spots, a selection was made of frauds that could be uncovered through data analysis. Several kinds of fraud fall beyond the scope of this investigation, there we know those kinds won't emerge out of the available data. The selected frauds are the so called known frauds. Aside from these known frauds, also unknown fraud will be aim of the study. In the following section the data engineering and the selected known frauds will be highlighted.

4.1 Data Engineering

The way data is looked at, organized and investigated is of primary interest in research using data mining. This is called data engineering. The main objective in this research is to detect fraud. But what particular aspect we are looking for is fraudulent?

Our data consists out of records. Such a record is described by attributes (date, person, value, account, movement...). In theory a record cannot be fraudulent. A set of records on the other hand, forming a transaction, can be fraudulent. Take for example a record that describes the payment of an invoice to supplier X. Another record describes the preceding purchasing order to X for a smaller amount of money than on the mentioned invoice. Separately, these two records aren't fraudulent. Only when combined into one transaction one can judge this transaction fraudulent. But are fraudulent transactions what we are looking for? In fact no. Like records constitute transactions, transactions constitute the behavior of a fraud. The frauds are the ultimate objects we are interested in.

The only way of discovering frauds is by investigating their behavior. Observing employees' behavior can take place by examining corresponding attributes. These attributes of employees are built on attributes of transactions, which in turn are built on attributes of records. For getting even better insights, new attributes are added to the ones already available. The attributes at the highest level are the base for a suspicion score. This score gives eventually an idea about the probability an employee is fraudulent.

An assumption made in this engineering is that the behavior of a fraud is significantly different from the behavior of an honest employee. If this is not the case, we will not find any differences between attributes describing the behavior of a fraud and the attributes describing the behavior of a regular employee. Hence no suspicion scores will be significantly different from the other scores.

4.2 Known Frauds

4.2.1 Double Payment of Invoices

The first known fraud selected is double payment of invoices. We restrict this fraud to cooperation between an employee and a supplier. The employee enters the invoice twice into the system. This has to happen under slightly changed circumstances, because the administration system prohibits to entry an exact copy of an invoice. Whenever the invoice number for example is changed a bit, this control is circumvented. After the doubled invoice is paid twice to the supplier, a kickback comes to the employee.

4.2.2 Changing Purchasing Order after Release

After the creation of a purchasing order, the internal system starts a work flow in order that two hierarchical authorized persons approve and release this order. After release, the order is printed and sent to the supplier this order was created for. If an employee, in charge of entering those orders into the system, changes the order afterwards, it has to be released again. However, if the changes are small enough, this isn't the case. Epsilon works strict percentages for judging what is 'small enough' to change an order without starting a new release strategy. Employees know these percentages, and can abuse them for personal enrichment, again in cooperation with the supplier.

4.2.3 2% Deviation of Purchasing Order

After sending a purchasing order to a supplier, the ordered goods and the accompanying invoice will be received. Since the approval has already taken place at the moment of the order creation, this hasn't have to occur again. For payment, the invoice is compared to the quantity of what is received (entered into the system at receival) and the price which was agreed on in the order. If there is a match between both these factors, the invoice will be paid. This match is checked systematically and leaves room for deviation, preventing an overload of work for minor adjustments. Adjustments that don't prohibit payment, must be smaller than 2%. This rule counts for every item line separated. As with the changing of a purchasing order after release, this information can be communicated to suppliers and a combine can be set up between an employee and supplier.

4.3 Unknown Frauds

Unknown frauds are all frauds possible at the procurement cycle which we aren't looking for explicitly. We aren't explicitly looking for detecting them, because we don't know how they would function, hence they are (for this moment) unknown to us. It is important to keep in mind that the detection of unknown fraud is part of the scope of investigation.

4.4 Data Selection

The data used for this study is originated from the Enterprise Resource Planning(ERP) system used by Epsilon. The data contains information about purchasing orders, goods receipts and invoices (who created it, who approved or released, how many items are ordered, for what price, which date, for whom, etc.). Aside from this, information about the flow a financial document follows is available.

4.5 Data Mining Techniques Selected

The data mining techniques selected will be clustering, outlier detection and sequence rules. If feasible, also Bolton and Hand (Bolton and Hand)'s peer group analysis will be used.

5 Current Status

For the moment, the authors are manipulating the data. Desired attributes of both invoices (or purchasing orders) and employees are being created. This is done for each kind of known fraud, so eventually we will have six data sets to work on (for starters). By the time of the Research Symposium, the attributes will be communicated and first analysis will be presented.

References

ACFE (2006). 2006 acfe report to the nation on occupational fraud and abuse. Technical report, Association of Certified Fraud Examiners.

Barson, P., S. Field, N. Davey, G. McAskie, and R. Frank. The detection of fraud in mobile phone networks. *Working paper University of Hertfordshire, Hatfield*.

Bologna, G. J. and R. J. Lindquist (1995). *Fraud Auditing and Forensic Accounting* (2 ed.). John Wiley & Sons.

Bolton, R. and D. Hand. Unsupervised profiling methods for fraud detection.

Bolton, R. and D. Hand (2002). Statistical fraud detection: A review. *Statistical Science 17*(3), 235–255.

Bonchi, F., F. Giannotti, G. Mainetto, and D. Pedreschi (1999). A classification-based methodology for planning audit strategies in fraud detection. In *KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA. ACM Press.

Bordoni, S., R. Emilia, and G. Facchinetti (2001). Insurance fraud evaluation - a fuzzy expert system. In *FUZZ-IEEE*, pp. 1491–1494.

Brause, R., T. Langsdorf, and M. Hepp (1999). Neural data mining for credit card fraud detection.

Burge, P. and J. Shawe-Taylor (2001). An unsupervised neural network approach to profiling the behavior of mobile phone users to use in fraud detection. *Journal of Parallel and Distributed Computing 61*, 915–925.

Cahill, M., D. Lambert, J. Pinheiro, and D. Sun (2000). Detecting fraud in the real world.

Cortes, C., D. Pregibon, and C. Volinsky (2002). Communities of interest. *Intelligent Data Analysis 6*, 211–219.

Cox, K., S. Eick, and G. Wills (1997). Viual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery 1*, 225–231.

Davia, H. R., P. C. Coggins, J. C. Wideman, and J. T. Kastantin (2000). *Accountant's Guide to Fraud Detection and Control* (2 ed.). John Wiley & Sons.

Deshmukh, A. and T. Tallur (1997). A rule based fuzzy reasoning system for assessing the risk of management fraud. *Journal of Intelligent Systems in Accounting, Finance & Management 7*(4), 669–673.

Estévez, P., C. Held, and C. Perez (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications 31*, 337–344.

Ezawa, K. J. and S. W. Norton (1996). Constructing bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert 11*(5), 45–51.

Fan, W. (2004). Systematic data selection to mine concept-drifting data streams. *Proceedings of SIGKDD04*, 128–137.

Fanning, K. and K. Cogger (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management 7*, 21–41.

Fawcett, T. and F. Provost (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery 1*(3), 291–316.

Fawcett, T. and F. Provost (1999). Activity monitoring: Noticing interesting changes in behavior. In Chaudhuri and Madigan (Eds.), *Proceedings on the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, pp. 53–62.

Green, B. and J. Choi (1997, Spring). Assessing the risk of management fraud through neural network technology. *Auditing 16*(1).

He, H., J. Wang, W. Graco, and S. Hawkins (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications 13*(4), 329–336.

Lin, J., M. Hwang, and J. Becker (2003). A fuzzy neural network for assissing the risk of fraudulent financial reporting. *Managerial Auditing Journal 18*(8), 657–665.

Maes, S., K. Tuyls, B. Vanschoenwinkel, and B. Manderick (2002). Credit card fraud detection using bayesian and neural networks. *Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies, January 6-19, 2002*.

Major, J. and D. Riedinger (2002). EFD: a hybrid knowledge/statistical-based system for the detection of fraud. *The Journal of Risk and Insurance 69*(3), 309–324.

Michalski, R. S., I. Bratko, and M. Kubat (1998). *Machine Learning and Data Mining - Methods and Applications.* John Wiley & Sons Ltd.

Murad, U. and G. Pinkas (1999). Unsupervised profiling for identifying superimposed fraud. *Lecure Notes in Computer Science 1704*, 251–262.

Pathak, J., N. Vidyarthi, and S. Summers (2003). A fuzzy-based algorithm for auditors to detect element of fraud in settled insurance claims. *Odette School of Business Administration Working Paper No. 03-9*.

Phua, C., V. Lee, K. Smith, and R. Gayler (2005). A comprehensive survey of data mining-based fraud detection research.

PWC (2005). Global economic crime survey 2005. Technical report, Price-WaterhouseCoopers.

Rosset, S., U. Murad, E. Neumann, Y. Idan, and G. Pinkas (1999). Discovery of fraud rules for telecommunications: Challenges and solutions. In *KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 409–413. ACM Press.

Shao, H., H. Zhao, and G. Chang (2002). Applying data mining to detect fraud behaviour in customs declaration. In *Proceedings of the First International Conference on Machine Learning and Cybernetics.*

Viaene, S., G. Dedene, and R. Derrig (2005). Auto claim fraud detection using bayesian learning neural networks. *Expert Systems with Applications 29*, 653–666.

Wasserman, S. and K. Faust (1998). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

Wells, J. T. (2005). *Principles of Fraud Examination.* John Wiley & Sons.

Witten, I. and E. Frank (2000). *Data mining: practical machine learning tools and techniques with Java implementations.* San Francisco, Calif.: Morgen Kaufmann.

Yang, W.-S. and S.-Y. Hwang (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications 31*, 56–68.