Data analysis can be a strategic weapon in your company's management and control of fraud

Fraud or scams — euphemistically called economic offenses — are a dominant white-collar crime in today's business environment. An unfortunate but rather well-known fact is that many businesses and government organizations, particularly in financial and related services, suffer from fraud of various kinds. Fraud bleeds businesses to the tune of hundreds of billions of dollars worldwide, annually. Continued prevalence of such malpractices on a large scale can have disastrous long-term consequences not only for the businesses involved but also for the investors, financial institutions, government, and economy, in general.

Today's highly automated business systems collect vast amounts of data regarding almost all kinds of business transactions and activities. With the advent of data warehousing and corporate memory systems, you can now access both current and historical business data. Clearly, evidence of fraud and fraudulent activities is partly hidden in these enormous quantities of data. Data analysis techniques can help businesses perform effective fraud management to prevent losses and bring the culprits to justice.

Fraud management involves a whole gamut of activities: early warnings and alarms; tell-tale symptoms and patterns of various types of fraud; profiles of users and activities; fraud detection, prevention, and avoidance; minimizing false alarms and avoiding customer dissatisfaction; estimating losses; risk analysis; surveillance and monitoring; security (of computers, data, networks, and physical facilities); data and records management; collection of evidence from data and other sources; reports; summaries; data visualization; links to management information systems and operation systems (such as billing and accounting); and control actions (such as prosecution, employee education and ethics programs, hotlines, and cooperation with partners and law enforcement agencies).
Several critical issues make building fraud management systems a challenging and difficult task: enormous volumes of data with complex structure; changing behavior of users, business activities, and fraudsters; continuous evolution of newer fraud particularly to bypass existing detection techniques; need for fast and accurate fraud detection without undue burden on business operations; risks or false alarms; and social issues such as privacy and discrimination.

In this article, I'll take a brief look at the various types of fraud and the means and processes, in particular software-based techniques, that you can use to detect, investigate, and prevent them.

**WHAT IS FRAUD?**

Oxford Advanced Learner's Dictionary defines fraud as "an act of deceiving illegally in order to make money or obtain goods." Indeed, in fraud, groups of unscrupulous ("morally challenged," if you prefer!) individuals manipulate or influence the activities of a target business with the intention of making money or obtaining goods through illegal or unfair means. Fraud cheats the target organization of its legitimate income and results in a loss of goods, money, and even good will and reputation. Fraud often employs illegal and always immoral or unfair means.

Outright criminal activities — typically involving violence or other physical means — such as break-ins, thefts, industrial espionage, sabotage, attacks and robberies, and so forth are usually excluded from the scope of fraud. But even within a particular organization, the full scope of what exactly constitutes fraud isn't always clear. A particular difficulty is distinguishing fraud from losses due to incompetence, procedural lapses, accidents, mismanagement, wrong decisions, or business risks. General economic offenses also include criminal acts other than fraud: money laundering, financing of criminal or antinational activities, corruption, bribery, kickbacks, and so on.

Nevertheless, due to their potential for significant negative impact, fraud has been studied in-depth as a phenomenon. Luckily, fraud falls into typical similar types that share common characteristics, means, and methods. Just as a garden-variety house theft can occur in only some specific ways — break-in, lock picking, gaining entry, and confidence by misrepresenting identity — fraud shares similar modus operandi. Consequently, an organization can take advantage of these commonalities to establish business practices to protect itself from fraud and resultant losses. Of course, any particular fraud in an organization need not meet all of the characteristics.

Fraud often consists of many instances or incidents involving repeated transgressions using the same method. Fraud instances can be similar in content and appearance but usually aren't identical.

Fraud investigations are complex, time-consuming, and tedious activity and require a great deal of knowledge of finance, economics, business practices, market analysis and business conditions, investigative skills, and law. A comprehensive investigative and surveillance business process for fraud management (often set up in the form of a fraud control center within an organization) often includes a number of steps, activities, and deliverables. I'll take a brief look at the core of this business process: data analysis.

**DATA ANALYSIS TECHNIQUES FOR FRAUD DETECTION**

The techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence (AI) techniques. Many commercial tools are
available for fraud detection that provide a variety of techniques from either of these areas, although usually not in any single integrated tool. Important statistical data analysis techniques for fraud detection are:

- Data preprocessing techniques for detection, validation, error correction, and filling up (estimation) of missing or incorrect data.
- Calculation of various statistical parameters such as averages, quartiles, performance metrics, probability distributions, and so on. For example, the averages may include average length of call, average number of calls per month (or per day), and average delays in bill payment.
- Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
- Computing user profiles (classifications of users, customers, and orders into various categories) and statistical characterization of these profiles (in terms of parameters, probability distributions, and so forth).
- Time-series analysis of time-dependent data.
- Clustering and classification to find patterns and associations among groups of data.
- Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.

In addition, a number of auxiliary tools can help surveillance personnel quickly grasp the nature of business data and activities. These include canned queries, summary reports, data visualization in various forms, software filters in the form of early warning indicators, alarm conditions, and so on. Usually, these techniques require considerable human expertise and active participation. Also, they're used in a sort of iterative way, where suspicious transactions are first identified and then further investigated to locate the victims, suspects, and their methods, which are then investigated to enable prevention or gather evidence.

I've already remarked that fraud management is a knowledge-intensive activity. Therefore, applications of knowledge-based techniques from AI are a natural idea. Important AI techniques used for fraud management include:

- Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
- Expert systems to encode expertise for detecting fraud in the form of rules.
- Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either automatically (unsupervised) or to match given inputs.
- Machine learning techniques to automatically identify characteristics of fraud.
- Neural networks that can learn suspicious patterns from samples and used later to detect them.

Other techniques such as Bayesian networks, decision theory, and sequence matching are also used for fraud detection.

MEDICAL FRAUD

I'll illustrate some of these techniques to handle the problem of fraud detection in a hypothetical and highly simplified medical insurance claims database. This database (as maintained by the insurance company and populated from the claim documents submitted by patients) consists of a single table and has the following format:

1. Patient ID (SSN)
2. Sex (M/F)
3. Age (0 to 120)
4. Address
5. Claim Date
6. Illness Category, Illness ID, and Illness Description (may be more than one illness)
7. Illness Duration Start Date – End Date
8. Hospital ID(s)
9. Doctor ID(s)
10. IDs of diagnostic tests performed
11. Names of medicines given
12. Other treatments (for example, physiotherapy)
13. Diagnostic tests bills
14. Medicine bills
15. Other treatment bills
16. Hospital bills
17. Doctors' charges
18. Misc. amount (all other costs)

For a summary of the most common characteristics of fraud and a list of common types of fraud, see the sidebars "General Characteristics of Fraud" and "Business Fraud" available online at www.intelligententerprise.com/020528/feat3_1.shtml.
Let's evaluate whether a new specific claim is "suspicious" in some way. If so, the claim can be processed in a different way — cancel claim payment, proceed with claim payment, recall claim, reduce payment amount, or seek clarification from hospital or patient.

For the purpose of evaluating a new claim, you can often define various criteria or indices for suspiciousness. For each criteria or index, the claim gets a score; typically, high-score values in a specific index indicate greater suspiciousness. Thus, a claim that has high scores for many criteria is more suspicious. Examples of such criteria include:

- The net amount is too large as compared to the average amount in similar claims.
- The cost of one or more diagnostic tests is too large as compared to the average amount in similar claims.
- The percentage of one diagnostic test costs to the net amount is too high as compared to the average percentage in similar claims.
- The previous two scores can be adapted for medicine costs, doctors' charges, hospital bills, and other costs.
- The claim is a duplicate (a very similar claim by the same patient was paid in recent past).
- The address of patient, hospital, or doctor is suspicious (missing ZIP Code, address includes P.O. Box number, errors in address components — incorrect phone number, ZIP Code, town name, or email address).

You can define many more such indices. All such indices have to be defined rigorously; the previous descriptions are merely indicative. Ideally, the fraud control system can provide a facility to dynamically define such indices outside the system so that enhancements are easily possible. Because the indices represent knowledge about the fraud detection in claims warranty data, a rule language can capture it in a knowledge base. The system can provide a facility that lists similar claims to the given claim (based on k-nearest-neighbor algorithms, for example), along with a similarity matching score. This facility would enable the end user to evaluate the given claim with respect to similar claims. From a pool of already known fraudulent claims, machine-learning algorithms can construct a classification (such as a decision tree) that can help evaluate a new claim.

As a simple example, you can check the disease (illness) ID against the duration and costs. Using the historical claims database, you can easily get a histogram of the hospital duration bins (0 to 2 days, 3 to 5 days, and so on) against the number of claims (this histogram will be for a specific illness ID, sex, and age group). You can then compare the claim duration against this histogram. If it falls in a sparsely populated bin, then it's at least a bit suspicious. Clustering of historical data can be used to automatically detect such outliers.

Several types of calculations can be performed for fraud detection, such as regression analysis and time-series analysis. In time-series analysis, the time-stamped data is analyzed for trends, seasonal patterns, and outliers. The series is first transformed, if necessary, so that the variance is constant. Additional assumptions may be needed because the observations in claims data aren't necessarily at regular time intervals. Several time series in the claims data can be analyzed using time-series analysis techniques. For example, the NO. OF CLAIMS and NET_AMOUNT (or any other component of claim amount) for any specific or all hospitals, for any specific or all illness IDs, and for any specific or all patients.

Suppose X is the time-varying quantity (NO. OF CLAIMS) in a time-series; let X(1) denote value of x at time 1. A graph can be used to plot week-to-week changes (X(I+1) - X(I)) in the time-varying quantity (NO. OF CLAIMS). This can be used to quickly identify the outliers. Another graph that plots week-to-week percentage change is X100 * X(I+1) / X(I) in the time-varying quantity. Autocorrelation and other techniques can be used to study these time series.

The following are some variables that are important for fraud detection in the claims data. Multiple regression analysis can be performed on chosen subsets of these variables:
• Age
• Sex
• Hospital ID
• Illness ID
• Duration of illness
• Various cost components
• Net amount.

Statistical analysis can also be performed for identifying outliers:
• Test cost outliers
• Hospital charges outliers
• Medicines cost outliers
• Illness duration outliers
• Combination outliers (doctors charges and net amount).

Some important temporal parameters for a claim include CLAIM_DATE, illness start and end dates (duration of illness). The difference between CLAIM_DATE and ILLNESS_START_DATE, called CLAIM_DELAY, is an important independent variable. The relationship between the two independent variables NET_AMOUNT (on X-axis) and duration of illness (on Y-axis) can be shown in a scatter plot (for only those claims for a specific illness ID). You may find, for example, that most claims above $20,000 have a long duration of illness. The Pearson correlation coefficient R for these two variables can be computed easily and indicates how closely related these two variables are. Analysis of variance can be used to check if the mean duration of illness is equal for, say, all hospitals. Moreover, these comparisons can be done for claims of different NET_AMOUNT bins. If not, further tests can be performed for ensuring that there's no special behavior by specific hospitals.

All such statistical analyses need to be studied in-depth and defined for the specific tasks of fraud detection and control in the medical claims domain. A large number of predefined statistical calculations oriented for detecting suspicious data can be provided.

Fraud is an important phenomenon in today's wired commercial world. Fraud causes huge losses and damages an organization's reputation and good will. Fraud management is a complex and knowledge-intensive process involving deployment and effective use of tools based on a plethora of statistical and AI techniques. The author wishes to thank Prof. Mathai Joseph for his support. Thanks to Dr. Manasee Pulbikan for her patience, hope, and confidence.

STRATEGIC KNOWLEDGE

(Continued from page 12)

they're being offered merely a repackaging of familiar BI tools? Or is it something that is — to borrow words from Intelligent Enterprise contributing editor Philip Russom — "interactive and analytic (not just static, like a report), [with] procedures that guide the user through analysis, the way other types of applications automate linear business processes." ("Nonanalytic Nonapplicatons," Decision Support, Feb. 1, 2002.)

Morris, in his presentation at IDC Directions, described how analytics would combine with content and collaboration services to form next-generation enterprise applications. Holding this package together will be increased "verticalization," which follows what Morris identified as a familiar evolutionary trend in software applications. After an initial phase of "horizontal" tool development, packaged suites form. Then, to create differentiation and higher value, the suites begin to encompass industry- or business function-specific expertise. Advanced enterprise portals, offering dynamic, role-based process management, will serve as the secure, interactive platform that could, in a sense, escalate the user's experience up to the appropriate level.

Where does this leave the BI vendors? Will they be able to compete effectively in both horizontal and vertical markets? Crystal Decisions is one vendor that could have a strong position, thanks in part to its relationships with Microsoft and SAP AG; the Microsoft .Net relationship will be particularly intriguing as BI Web services evolve. Crystal is also developing more vertically oriented analytic applications. At perhaps the other end of the spectrum, SAS Institute is investing heavily in building vertical domain expertise to fit with the operations research capabilities of its software tools. Biotechnology, demand planning for retail, and supplier relationship management are three key areas for SAS analytic applications.

ADIEU, CFO

As XML and federated data management approaches mature, BI will float free of its moorings within the traditional data warehousing context to explore new analytic waters. Within the "intelligent" enterprise applications that Morris describes, BI tools will have to negotiate data streams containing structured and unstructured data. That's because so many of these applications will have new masters: that is, C-level figures who don't necessarily see the world through the filter of accounting. Instead, they'll want to analyze profitability and return on investment based on specific business processes, marketing campaigns, or other activity — using approaches such as balanced scorecard frameworks and activity-based costing.

Our recent recession has taught line-of-business managers that they need new instruments; otherwise, they're at a serious disadvantage vs. competitors who have sophisticated analytics that can look beyond what traditional, accounting-oriented BI vendors can dance to a different rhythm.