Statistical Methods for Fighting Financial Crimes

Agus SUDJIANTO

Bank of America
Charlotte, NC 28255
(*Agus.Sudjianto@bankofamerica.com*)

Sheela NAIR

Bank of America
Charlotte, NC 28255
(*Sheela.nair@bankofamerica.com*)

Ming YUAN

Georgia Institute of Technology
Atlanta, GA 30332
(*myuan@isye.gatech.edu*)

Aijun ZHANG

Bank of America
Charlotte, NC 28255
(*Aijun.zhang@bankofamerica.com*)

Daniel KERN

Bank of America
Charlotte, NC 28255
(*Daniel.c.kern@bankofamerica.com*)

Fernando CELA-DÍAZ

Bank of America
Charlotte, NC 28255
(*Fernando.cela-diaz@bankofamerica.com*)

Financial crimes affect millions of people every year and financial institutions must employ methods to protect themselves and their customers. The use of statistical methods to address these problems faces many challenges. Financial crimes are rare events that lead to extreme class imbalances. Criminals deliberately attempt to conceal the nature of their actions and quickly change their strategies over time, resulting in class overlap and concept drift. In some cases, legal constraints and investigation delays make it impossible to actually verify suspected crimes in a timely manner, resulting in class mislabeling or unknown labels. In addition, the volume and complexity of financial data require algorithms to be not only effective, but also efficiently trained and executed. This article focuses on two important types of financial crimes: fraud and money laundering. It discusses some of the traditional statistical techniques that have been applied as well as more recent machine learning and data mining algorithms. The goal of the article is to introduce the subject and to provide a survey of broad classes of methodologies accompanied by selected illustrative examples.

KEY WORDS: Anomaly detection; Classification; Fraud detection; Machine learning; Money laundering.

1. INTRODUCTION

Financial crimes refer to a broad category of crimes against property, committed by individuals and organizations to obtain a personal or business advantage. As described by the Federal Bureau of Investigation (2005), they are characterized by deceit, concealment, or violation of trust, and are not dependent upon the application or threat of physical force or violence. Examples include money laundering, credit/debit card fraud, embezzlement, counterfeiting, mortgage fraud, and insider trading, to name a few. These crimes cost several billions of dollars a year and affect the lives of millions of people. By law, U.S. financial institutions are required to carry a substantial amount of the responsibility for combating financial crimes. Of course, these institutions must also protect their customers and shareholders from financial loss. This has led to extensive research and the development of detection systems.

This article focuses on two specific types of financial crimes that pose large threats: *money laundering and retail banking fraud*. In money laundering, the criminals hide the true origin of funds by sending them through a series of seemingly legitimate transactions. The main purpose of laundering money is to conceal the fact that funds were acquired as a result of some form of criminal activity. These laundered funds may, in turn, be used to foster further illegal activities such as the financing of terrorist activity, trafficking of illegal drugs, support of prostitution rings, or smuggling of weapons. Even the laundering of legitimate funds to avoid reporting them to the government (e.g., tax evasion) leads to substantial costs for society. The U.S. Internal Revenue Service estimates over $300 billion in taxes went unpaid in 2001 alone.

In retail banking fraud, the criminal attempts to achieve financial gain at the expense of legitimate customers or financial institutions through any retail banking transaction channel, such as credit cards, debit cards/ATM's, online banking, or checks. Most of the fraud detection research focused on credit card fraud, which was estimated at close to $1 billion in the U.S. and $10 billion worldwide (Ghosh and Reilly 1994; Aleskerov, Freisleben, and Rao 1997). This poses a serious problem for financial institutions as they increasingly assume responsibility for all unauthorized transactions. Debit card fraud is also a growing concern as these cards gain volume share; debit/ATM fraud losses in the U.S. were estimated at $2.75 billion (Gartner 2005). A common pattern, *skimming*, involves collecting personal identification numbers (PIN) from compromised point of sale readers or ATM's and cloning the card's magnetic strip. This enables criminals to obtain cash without any human contact in a way that is particularly difficult to trace. The financial

TECHNOMETRICS, FEBRUARY 2010, VOL. 52, NO. 1
DOI 10.1198/TECH.2010.07032

impact of this type of fraud has increased as a result of relaxed overdraft policies at many banks, which allow transactions to be made on an account with insufficient funds. Although the laws protecting customers from fraud are weaker for debit cards than credit cards, the financial institution will still cover the customer's loss in most cases.

Online-banking fraud losses were estimated at $2 billion in 2004 (Montana 2006), mostly driven by identity theft: criminals obtain personal information such as the name, address, and social security number of a legitimate customer. This can be done by breaking into a financial institution's or a customer's computer system or by impersonating online banking websites (*phishing*), automated call centers (*vishing*), or a financial institution or government employees (*social engineering*). The information is then used to take control of the customer's assets, to make unauthorized purchases, or to open new lines of credit in the name of the customer. Even when the charges are paid by the financial institution, identity theft can be a tremendous burden on the victim in the form of a ruined credit background that can lead to a long and difficult repair process. In more refined schemes, a criminal persuades a legitimate customer to act as an intermediary on a seemingly legal transaction. The customer receives a transfer in his bank account, retains a fraction as compensation, and forwards the rest to a third account; what the customer ignores is that the origin of the transfer is a bank account *phished* online. By acting as an intermediary, the legitimate customer has involuntarily become an accessory to the crime.

Check and deposit fraud account for about $677 million in losses per year (Montana 2006). In this case, a criminal exchanges a worthless item for cash—for example, a fake check or a check drawn on an account with insufficient funds. Classic *scams* persuade legitimate customers to accept a seemingly legitimate check in exchange for a smaller amount of cash than the amount shown on the check; when the check bounces, the customer loses the full amount.

The criminal strategies used in money laundering and retail banking fraud are inherently different, and therefore, need to be fought using different methods. Retail banking fraud usually involves a few transactions occurring over a short period of time on relatively new accounts. The strategy is usually to commit the fraud as quickly as possible, before it is noticed by the customer or by the bank, and then move on to another victim. While the amount involved in a single incident is small, large gains are accumulated over several attempts in a short period of time. In contrast, a typical money launderer will try to maintain lengthy and healthy relationships with established financial institutions because time and patience are needed when trying to move large amounts of money through the financial system.

Probably the biggest difference between retail banking fraud and money laundering is that a financial institution rarely finds out if a money laundering suspect was actually guilty of the crime. Typically, as soon as the case is detected, the information is turned over to the government for follow up, and the financial institution is unaware of the end result of subsequent investigations and prosecution. In the case of fraud, however, the financial institution does have access to such information, and therefore, detection systems can be trained with reliable data.

Finally, there are many different types and scales of retail banking fraud and money laundering, so it is difficult to solve the problem with a one-size-fits-all approach. For example, different methods are required to detect a small company trying to avoid taxes versus a large criminal organization laundering vast amounts of money from the sale of illegal drugs.

2. OUTLINE

The goal of this article is to provide an introduction to the topic, an overview and examples of selected methods in the literature, and practices for detecting financial crimes. In doing so, we also present some new techniques to alleviate problems encountered in practice. There was extensive research on detecting financial crimes using traditional statistical methods as well as more recent machine learning techniques. See, for example, Bolton and Hand (2002), who provided a comprehensive review of statistical methods used in credit card, telecommunications and medical fraud detection, intrusion detection, and money laundering.

Most of the attention in the literature focuses on fraud, particularly on credit card fraud; fewer reports exist on money laundering. In order to provide a complementary view, we include more examples on the latter topic. While there is some overlap among the various statistical methods, we group them into two broad classes as supervised learning and unsupervised learning (anomaly detection). The performance of the different techniques is very data-dependent and we provide some illustrative performance comparisons. Throughout the article, we intentionally simplify the examples by considerably reducing the number of features or variables involved in the models as well as their detailed explanation. We recognize that this simplification has a significant drawback because it may make the examples look superficial at a first glance. But we have to do this due to the confidential nature of the problem and the need to keep the presentation clear.

Section 4 provides a review and examples of supervised learning methods used for profiling and classification. Several popular machine learning algorithms and their applications are discussed and illustrated with examples. Section 5 provides a review and examples of unsupervised learning methods. A more in-depth discussion is provided for the anomaly detection technique, with an example and enhancement of existing methodology. The article ends with some concluding remarks.

3. DETECTION STRATEGIES AND STATISTICAL CHALLENGES

In our context, *detection* is simply the ability to discover that a financial crime occurred. A detection system will try to identify patterns and trends of suspicious behavior. Typically, the system will generate a *suspicion score* indicating how likely a case is to be criminal. Cases exceeding a certain suspicion-score threshold will be investigated. The effectiveness of the detection system will ultimately depend on the speed at which the crime is detected, the range of crimes that can be detected, and the number of false alarms generated.

The detection process starts when a customer applies for an account and the goal is to identify criminals before the account

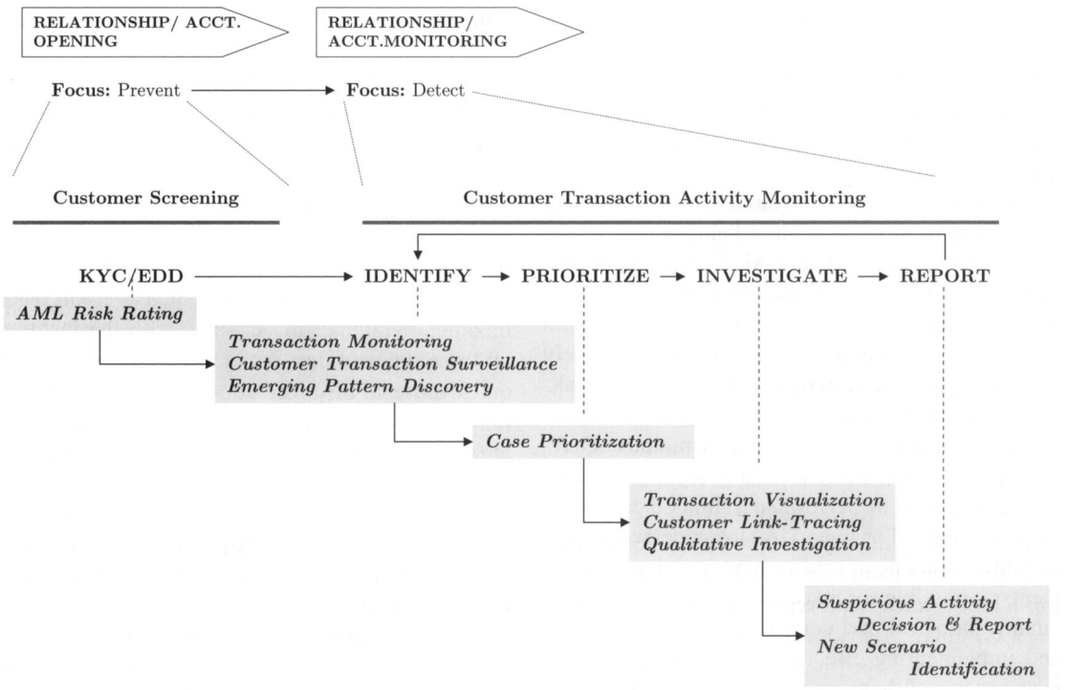

Figure 1. General AML process.

is opened. The focus then switches to monitoring the activity of the new account. Figure 1 depicts an overview of the process of antimoney laundering (AML) detection. During account opening, an institution is required to enforce a *Know Your Customer* (KYC) policy and perform an *Enhanced Due Diligence* (EDD) check to verify the identity of the customer. At this stage, a *risk rating* model (which can be either statistical or rule-based) is usually applied to score the customer's potential risk. Once customers are on-board and begin to conduct transactions, the next phase of transaction activity monitoring is continuously applied. There are typically four stages of monitoring:

- *Suspicious Activity Identification*. This step may employ various strategies such as monitoring all transactions based on historically known suspicious patterns, surveillance of a high-risk customer's (identified by risk rating at account opening) transactions, or exploratory analysis to identify emerging patterns.
- *Case Prioritization*. A prioritization model is used to rank the many detected potential suspicious activities for further investigation. This step is typically required because of *false alarms* generated by the identification step, limited available resources to handle these cases, and varying degrees of criticality from case to case.
- *Case Investigation*. This is a human resource intensive step where experts investigate the case. A typical investigative process includes the following steps: retrieve transaction data of suspicious customers, separate *inflow* and *outflow* of funds, filter relevant transaction types (e.g., wire transfer, cash) and suppress irrelevant information, create summary statistics (e.g., daily, weekly) to extract features (e.g., total amount, frequency, average, maximum), gather related information (e.g., customer profile, income) from various sources (e.g., internet, third party database), and use heuristics and experience to *create a story*—Who is the customer? What banking products does the person use (e.g., checking, credit card, investment)? What kind of transactions does the person conduct (e.g., ACH, wire transfer, cash)? Which transaction channel does the person use (e.g., ATM, internet)? Where were the transactions conducted (e.g., geographic location, merchant)? What are the amounts and frequency? Were there any related incidents in the past?

  Data visualization tools are applied in dealing with a vast amount of transaction data. Link analysis techniques are applied to identify potential networks of involved entities by tracing transaction flows from various entities. Various sources of information, both quantitative (e.g., transaction data) and qualitative (e.g., internet search) in nature are used and the investigators apply their heuristics and qualitative judgment throughout the investigation process. Because this step is labor intensive, it is important to have highly effective identification and prioritization models in place prior to the investigation step.
- *Suspicious Case Reporting*. Once a case is deemed suspicious, the institution is required by law to file a report to the Financial Crimes Enforcement Network (FinCEN) under the U.S. Treasury Department. A well-constructed report is needed because law enforcement agencies rely on this report to conduct a further investigation. Additionally, this step provides a feedback loop of continuous learning to update the scenarios to identify future suspicious activities.

There are four broad categories of detection methods. The first class of methods monitors all transactions on an ongoing,

short-term basis and uses rules to flag suspicious incidents for investigation. These rules are usually simple since they are applied to very large datasets. A second approach analyzes data over a longer period of time using more complex statistical tools. This is typically done on smaller, summarized datasets so that the analyses and computations are manageable. This type of detection is usually done at an individual account level and can be used to determine if an account's behavior is consistent with its own history and with that of similar accounts. The third class of detection methods uses some criteria to prioritize a group of accounts based on how likely they are to contain suspicious activity. This list is then ranked from the most likely to the least likely in terms of crime potential and the top ranking activities are then investigated. The final class of methods searches for activity that is similar to known suspicious behavior. For example, if a person is involved in activity that is found to be suspicious, a query can be performed to find other activity where the person is involved or linked. A strong detection program utilizes all four methods to some extent in order to detect as many incidents as possible for investigation.

There are many statistical and computational challenges in developing and implementing effective systems for detecting financial crimes, but five of them are particularly important and will be discussed in greater detail: (1) volume and complexity of data, (2) class imbalance, (3) concept drift, (4) class overlap, and (5) class mislabeling.

3.1 Volume and Complexity of Data

A large financial institution has millions of customers and experiences thousands of customer transactions per second over multiple channels (e.g., internet, telephone, ATM, branch offices). This results in extremely large databases, often distributed across multiple, heterogeneous legacy systems. Often, decisions must be made in real time—a credit card fraud detection system is only useful if it can identify and stop a fraudulent transaction immediately. This places severe constraints on the computational complexity of the algorithms that can be feasibly implemented in the system.

Ideally, detection systems should work on data at three levels: transaction, account, and customer-levels. Transaction-level data include information about the current transaction (see Figure 2) such as amount, date, time, or location; account-level data includes information about the account history with the bank, such as average balance or account age; finally, customer-level data includes information such as credit risk scores, stated income, or number of accounts with the institution. Utilizing all three types of data is not always possible, as it will require quick retrieval of information from very large databases holding all account-holders' records. As a result, algorithms aimed at real-time detection will often need to make a decision solely based on the data present in the current transaction, and at most, a very limited amount of cached customer information.

Even when a real-time decision is not necessary, transaction-level, account-level, and customer-level data are high-dimensional and vary over time so that it is necessary to extract features that summarize the customer's transaction history. These can include the frequency of types of transactions, the average amount of transactions, the total amount of transactions, and so on. In the case of money laundering, the company must look for customers who are suspicious, and therefore, it is also necessary to use features that expose activities that are known to be common among criminals.

3.2 Class Imbalance

Financial crimes are rare events, resulting in severe class imbalance: the number of truly criminal cases or transactions is very small compared to the number of legitimate ones. Around 0.05%–0.1% of the 700,000 wires a day in the United States involved money laundering (OTA-ITC-630 1995). Dorronsoro

Acct.#	D/C	PostDate	TransAmt	TranCode	Description
999999999	D	1/24/2006	$1,295.00	1051	01051 CHECK CHECK
999999999	D	5/19/2005	$1,020.00	1051	01051 CHECK CHECK
999999999	D	1/24/2006	$10,000.00	1051	01051 CHECK CHECK
999999999	D	3/2/2005	$5.00	1513	01513 RETURNED ITEM CHARGE RETURNED ITEM CHARGE
999999999	D	2/24/2005	$5.00	1513	01513 RETURNED ITEM CHARGE RETURNED ITEM CHARGE
999999999	D	10/12/2005	$34.00	1203	01203 OVERDRAFT CHARGE OVERDRAFT CHARGE
999999999	D	7/13/2005	$60.00	1659	01659 CHECK CARD PURCHASE DR JM LAYTON AND EP LAY51
999999999	D	6/10/2005	$129.36	1105	01105 POS WITHDRAWAL COSTCO WHSE #0001 8442627516
999999999	D	6/14/2005	$51.49	1105	01105 POS WITHDRAWAL BED, BATH & BEYO 844262751650
999999999	D	6/10/2005	$168.44	1105	01105 POS WITHDRAWAL COSTCO WHSE #0001 8442627516
999999999	D	7/18/2005	$34.84	1105	01105 POS WITHDRAWAL COSTCO WHSE #0001 8442627519
999999999	D	5/24/2005	$33.20	1105	01105 POS WITHDRAWAL COSTCO GAS #00662 8442627514
999999999	D	6/22/2005	$158.65	1105	01105 POS WITHDRAWAL BED, BATH & BEYO 844262751730
999999999	D	6/10/2005	$190.64	1105	01105 POS WITHDRAWAL COSTCO WHSE #0001 8442627516
999999999	C	1/14/2005	$100.00	1003	01003 DEPOSIT DEPOSIT
999999999	C	8/9/2005	$20.00	1003	01003 DEPOSIT DEPOSIT
999999999	C	5/11/2005	$10,000.00	1003	01003 DEPOSIT DEPOSIT
999999999	C	8/31/2005	$3,300.00	1003	01003 DEPOSIT DEPOSIT 0831CA319P007160134679
999999999	C	6/29/2005	$2,079.95	1003	01003 DEPOSIT DEPOSIT
999999999	C	10/6/2005	$2,500.00	1003	01003 DEPOSIT DEPOSIT
999999999	C	1/30/2006	$22.43	1691	01691 AUTOMATIC DEPOSIT DEPOSIT MERCHANT BANKCD 26
999999999	C	1/30/2006	$22.43	1691	01691 AUTOMATIC DEPOSIT DEPOSIT MERCHANT BANKCD 26
999999999	C	6/16/2005	$64.97	1660	01660 REVERSE CHECK CARD PURCHASE THE HOME DEPOT 4
999999999	C	7/21/2005	$151.61	1660	01660 REVERSE CHECK CARD PURCHASE HARDWARE SALES

Figure 2. A sample of transaction activities from an account.

et al. (1997) reported a figure of 0.01% for credit card fraud in the Minerva system, which used to process more than 60% of all the Visa traffic generated in Spain, as well as all the foreign operations of Visa credit cards issued by Spanish credit institutions. Under such conditions, a trivial classifier that predicts all cases to be legitimate will have a very high success rate.

The classification of rare events is a difficult, but well-studied problem that arises in many different areas. A simple solution often used to balance the highly skewed class distributions is to subsample the majority class. More elaborate techniques exist, such as MetaCost (Domingos 1999) or synthetic minority oversampling (Chawla et al. 2002), but ultimately severe class imbalance imposes fundamental limits to classifier performance. These limits can only be overcome by changing the definition of the classification problem in a way that reduces the imbalance (Drummond and Holte 2005). For example, Fawcett and Provost (1997) changed the definition of a classification objective of a telecommunications fraud detection system from "phone calls" to "account-days." Dorronsoro et al. (1997) applied a parametric segmentation to input cases before classification in a credit card fraud detection problem.

In addition to the uneven classes, there are also different costs for the different types of misclassifications. A false positive, or incorrectly identifying a legitimate case as criminal, costs only as much as the resources spent investigating the case. However, failing to identify a truly criminal case (a false negative) can potentially cost the company a significant amount of money. Cost-sensitive methods are needed to address this issue.

3.3 Concept Drift

One of the main goals of detection systems is to identify general patterns of suspicious behavior. But even the formulation of this problem presents a challenge as these patterns are very dynamic and continuously evolve over time to bypass existing detection methods. Models must be continuously validated and adapted to accommodate these changing distributions and patterns.

The most basic method uses a fixed temporal window and retrains the algorithms at specific points in time. Widmer and Kubat (1996) proposed floating rough approximation (FLORA), a framework for learning and forgetting observations based on rough set theory that includes a window adjustment heuristic. Alternatively, the algorithm can be continuously retrained by reweighting the data so that more importance is given to recent observations (Klinkenberg and Uping 2002) or separate classifiers can be trained on different chunks of data and then combined as an ensemble (Street and Kim 2001; Wang et al. 2003).

If labels are available for new data, classification errors can be monitored to trigger retraining. The drift detection method (DDM) (Gama et al. 2004) monitors the error rate on classifier output. The early drift detection method (EDDM) (Baena-García et al. 2006) extends this concept to detecting gradual concept drifts by examining the distribution of distances between classification errors instead of the error rate.

A comprehensive survey of methods for learning from data under concept drift was given by Bifet and Kirkby (2009).

3.4 Class Overlap

Another challenge stems from the fact that criminals often try to conceal their activities by making illegal transactions seem as "normal" as possible, resulting in a substantial overlap between the delinquent and nondelinquent classes. It is common to have two transactions with similar characteristics, one of which is legal while the other is actually criminal. This is particularly the case in money laundering, where high-frequency deposit and withdrawal activities and the amount of deposits and withdrawals for a given day can be almost equal.

3.5 Class Mislabeling

It is not always possible to verify every case that was detected as suspicious. The most extreme case is money laundering, where the verification is conducted externally and the financial institution rarely learns of the final outcome of the investigation. This poses an important problem for machine learning, as detection algorithms will be trained with potentially mislabeled data and motivates the need for robust methods that can handle mislabeling effectively.

4. SUPERVISED LEARNING METHODS

Supervised learning methods use prior information on class membership and attempt to define a relationship between the set of inputs and outputs. In the case of fraud detection, this amounts to learning patterns of criminal and legal behavior to determine whether new activity is fraudulent. We describe three broad types of supervised learning techniques in the literature: supervised profiling, classification, and link analysis.

4.1 Supervised Profiling

If a database of tagged transactions or cases is available, *profiles* or distributions of relevant variables can be constructed for legitimate customer behavior and criminal behavior. Incoming transactions can then be automatically flagged for inspection on the basis of their similarity to criminal behavior, dissimilarity to expected legitimate behavior, or a combination of both.

In general, one profile of expected legitimate behavior is maintained per customer and one profile of fraudulent behavior is maintained per type of fraud. New transactions are compared with the customer's profile of legitimate behavior and with the different profiles of fraud. Deviations from expected behavior or similarity to known patterns of fraud may be a sign of criminal activity. In practice, these two criteria are usually blended into a single metric of suspiciousness via a Bayes ratio-related metric such as the weight of evidence (WOE). Given an observation of a vector of characteristics X of a customer with a profile ζ_i, the weight of evidence against a fraud profile φ_j is defined as

$$WOE = 10\log\left(\frac{P(X|\zeta_i)}{P(X|\varphi_j)}\right). \tag{1}$$

Used in this way, the WOE provides a measure of how much the customer's profile of legitimate behavior explains observed evidence compared to a profile of fraudulent behavior. As a rule of thumb, values lower than five indicate that the fraud profile

provides a better explanation for the observed behavior than the profile of legitimate customer behavior. See Siddiqi (2005) for a practical introduction to applications of WOE in the financial industry, with a focus on customer credit risk scorecards.

Rule-based profiles are a popular alternative to nonparametric estimates of the distributions of the variables of interest. The profile is summarized as a set of rules, and transactions are then matched to each rule. Rules can be defined from human experience, e.g., "any individual exceeding one million dollars worth of transactions in a single day shall be considered suspicious," or learned from data with a rule discovery algorithm. A rule-based approach has the advantage of being easy to implement on enterprise systems and easy to understand and interpret. Most importantly, their simple structure enables decisions on large amounts of (often streaming) data to be made very quickly. Alternatively, rule-based profiling can be used as a complement to more sophisticated analysis techniques to filter out transactions that are deemed to be safe. This is the method used by many credit card companies (see for example Chan, Fan, and Prodromidis 1999).

Profiles must be updated to reflect the dynamic patterns of criminal activity as well as changes in legitimate user behavior. This presents a challenge for static rule-based methods that are learned off-line, as they must be frequently validated and retrained. Wang et al. (2003) proposed using weighted rule ensembles as an alternative to full periodic retraining. A series of different rule sets are trained from sequential chunks of data. Weights are then assigned based on expected prediction accuracy on current test examples.

An additional problem is that the cost of false positives is dependent on the volume of criminal activity and the size of the workforce investigating the alarms, both of which constantly fluctuate. The frequency of detection must be kept under reasonable margins. This problem is often solved in practice by tuning thresholds in the rules manually. However, recursive or adaptive techniques that adjust rule parameters and thresholds as new information arrives can also be applied (Duda, Hart, and Stork 2000). The thresholds will be set based on training and historical data and will then self-adjust based on the streaming information. For example, a rule concerning the amount of transaction activity in a month can be employed with an initial threshold of $1,000,000. This may have an expectation of flagging approximately 1% of the customers. As the data stream is analyzed, it may become clear that a much lower percentage (say 0.2%) of customers actually exceed the threshold. The threshold will be lowered to capture the desired 1% of customers.

A pure recursive approach may not be feasible if the dataset is too large or information is received at too high a rate. In these circumstances, sampling can be employed to make the threshold calculations less computationally expensive. Depending on the specific rule, one can use various sampling techniques, particularly stratified sampling to sample within geographical regions or specific periods of time (e.g., the last business day of the month).

Profiling received a lot of attention in telecommunications fraud detection research, where considerable effort was made to overcome the weaknesses of standard rule-based profiling. These methods generalize easily to financial fraud applications and so are worth mentioning briefly. Fawcett and Provost (1997) proposed an adaptive user profiling method that uses user-specific rules and thresholds in detection. Cortes and Pregibon (2001) and Cahill et al. (2002) used the concept of an "account signature" for profiling accounts using streaming data: the user's behaviors are summarized using a set of features, which are described by a multivariate probability distribution. The signature is defined as a nonparametric estimate of the distribution. This estimate is updated sequentially, giving the highest weight to the most recent activity.

4.2 Classification

Both generative classifiers (e.g., hidden Markov models, Bayesian networks) and discriminative classifiers (e.g., support vector machines) were explored. The common goal of each technique is to use labeled data to train a model that determines the probability of each observation of being criminal.

Traditional statistical discrimination techniques such as logistic regression and linear discriminant analysis make use of linear decision boundaries for classification. These methods were applied for fraud detection (Mercer 1990; Foster and Stine 2004). We discuss in the following some more recent, nonlinear classifiers.

4.2.1 Support Vector Machines. Support vector machines (SVM's) (Vapnik 2000) work with a larger, transformed version of the feature space and find a maximum margin hyperplane that separates two classes of data. SVM's do well in classifying nonlinear separable groups; they do not require large training datasets and training converges to a unique global solution. These characteristics make SVM's attractive in problems such as credit card application fraud (Chen et al. 2004, 2005) and AML (Tang and Yin 2005). However, they are computationally intensive, the results are not easily interpretable, and many parameters of the algorithm must be specified, e.g., the type of kernel function used to transform the feature space and all its parameters. In practice, heuristics exist for selecting some of the parameters (Caputo et al. 2002; Hsu, Chang, and Lin 2009). Alternatively, general purpose search methods on parameter space such as the wrapper (Kohavi 1997) can be used. The search can be extended to feature space as well at the expense of computational cost. For example, Ahn, Lee, and Kim (2006) used genetic algorithms to optimize feature selection, instance selection, and kernel parameters simultaneously in a problem of bankruptcy detection.

4.2.2 Classification Trees and Ensemble Learning. Classification trees break the training data using different potential splits on every feature and recursively choose to split the data into two parts that minimize some measure of class impurity in the resulting subsets. Typical metrics of impurity include entropy, the Gini index, or classification error. Some methods can handle nonbinary splits as well. Three methods extensively used are CHAID (Kass 1980), CART (Breiman et al. 1983), and C4.5 (Quinlan 1993), as well as its successor, C5.0.

Decision trees can handle mixed numeric and categorical features and naturally accommodate nonlinear and nonsmooth decision boundaries and interactions between input variables. They also perform automatic feature selection by using only the features with the strongest classification power, although some limitations exist in practice (Witten and Frank 2005). They are

easy to interpret and their simple structure enables decisions on large amounts of streaming data to be made very quickly.

There are, however, disadvantages of tree-based methods. They are known to be unstable because of their hierarchical structure—small changes in the training dataset can generate very different trees and feature selection can be biased. Complex structures are needed to learn simple decision boundaries and simple asymmetric rules. For example, linear decision boundaries are approximated as staircase functions, thereby enforcing a hierarchical representation of the rules. Tree algorithms are also criticized for not generalizing well and pruning algorithms are usually needed to avoid overfitting. Conditional inference trees (Hothorn, Hornik, and Zeileis 2006) are a recent alternative that addresses the problem of selection bias and overfitting via permutation tests of association between the covariates and the target.

Tree ensembles attempt to overcome the limitations of simple tree methods by combining the outcome of multiple models in a single classification decision. From the standpoint of fraud detection, two leading ensemble learning methods are commonly used: random forest (Breiman 2001) and boosting (Freund and Schapire 1997).

Decision trees were extensively used in credit card fraud detection, especially in the field of credit risk scoring. Chan, Fan, and Prodromidis (1999) proposed AdaCost, a variant of the boosting algorithm AdaBoost especially suited for credit card transaction fraud detection. Carter and Catlett (1987) and Li et al. (2004) presented applications of ID3, a precursor to C4.5, to credit risk scoring. Lee and Chen (2003) compared CART and multivariate adaptive regression splines (MARS) to discriminant analysis, logistic regression, and neural networks for credit risk scoring.

Table 1. Classification errors, AML dataset

Method	Training error	Testing error
Logistic regression	8.9%	9.9%
CART	14.3%	14.4%
C4.5	1.8 %	13.5%
Random forest	0.0%	10.8%
Boosting (AdaBoost)	2.7%	10.0%
Boosting (LogitBoost)	0.0%	9.1%

To illustrate an application of decision trees, consider a money laundering detection case where three features are used: x_1: peer comparison score in terms of transaction volume, x_2: transaction speed score, and x_3: individual expected level of activity score. Figure 3 shows the scatterplots of the data. Due to confidentiality issues, we do not provide details of how the scores were developed. The example is intended to just illustrate the application of the techniques.

Table 1 summarizes the results of applying logistic regression, CART, C4.5, random forest, and two boosting methods. The data are split into equal sample sizes for training and testing purposes. The classification errors for both training and testing datasets are presented. In this particular case, simple decision trees exhibit poorer performance than logistic regression; how-

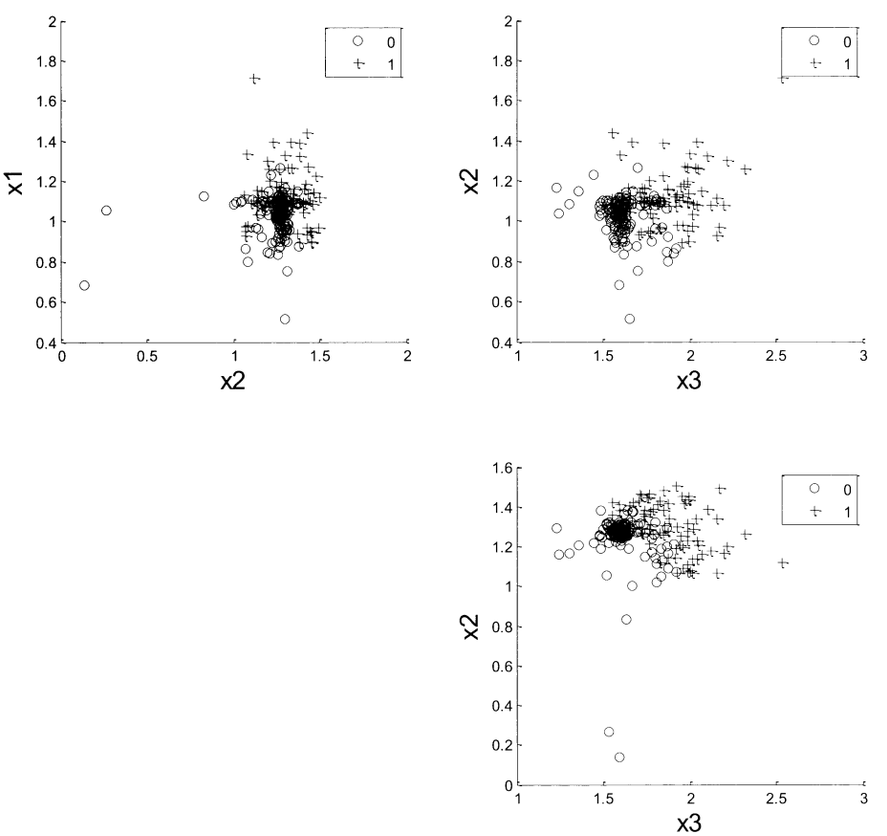

Figure 3. Scatterplots of money laundering detection data, 0: not suspicious and 1: suspicious.

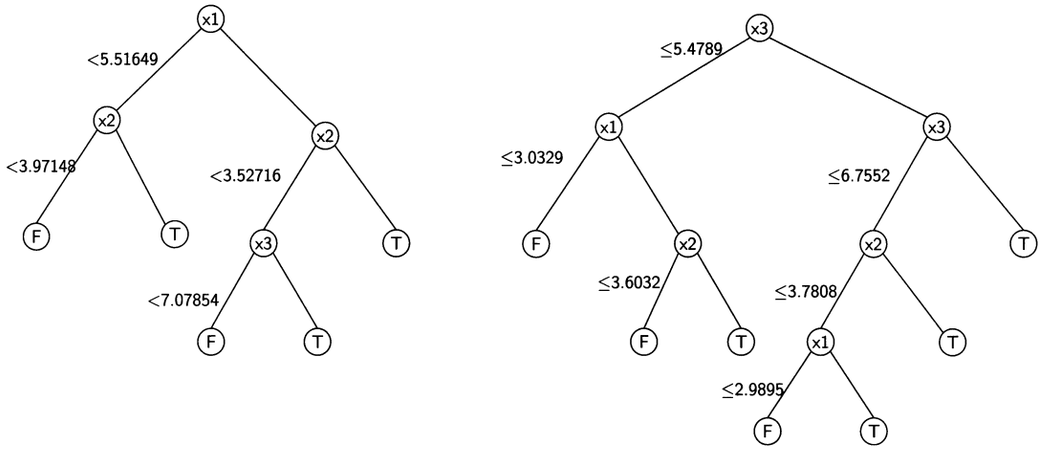

Figure 4. Decision trees of money laundering suspicious activity. CART (left), C4.5 (right).

ever, a logit-boosted tree with just 20 boosted iterations performs the best.

Figure 4 shows the trees produced by CART and C4.5. Interpretation is straightforward for both of them: for example, the tree on the left (CART) will flag transactions that exhibit low volume compared to peers and high speed, transaction with both high volume and high speed, as well as transactions with high volume, low speed, and a level of activity that is too high for what will be expected for that particular type of business.

Boosting algorithms can suffer overfitting in the presence of mislabeled classes (Opitz and Maclin 1999), whereas random forest is considered to be more robust to mislabeling than regular tree-based algorithms. Theoretically, random forest may experience degraded classification performance if a very large fraction of the feature space is irrelevant. In this case, the algorithm will randomly force these irrelevant features into the decision nodes (Rogers and Gunn 2005). In practice, such a level of variable irrelevance is uncommon. For illustrative purposes, we gathered a very large dataset for fraudulent transaction detection with 387 raw input variables. The dataset was obtained directly from a production system. No attribute selection was performed to prescreen irrelevant variables, nor was any data cleaning conducted to fix the data—particularly to remove potentially mislabeled data. The ratio of fraudulent transactions to legitimate transactions was 1 : 45 in the training dataset and 10% of the dataset was held out for testing. A random forest with 100 trees and 10 selection variables yielded 0.0%/27.0% false negative/false positive rates, despite potential irrelevant variables being present. Detailed inspection of the training dataset revealed that 83 variables provided little information about the outcome, but only 50 of those can be safely dismissed as completely irrelevant. Using 10 selection variables, it is highly likely that at least a few significant variables will make it into each decision node in random forest.

4.2.3 Classification Rules and Rule Ensembles. A ruleset is a set of logical tests that check whether an observation belongs to a class. In general, these sets are disjunctive (a logical OR) and each rule in the ruleset is expressed as a series of logical conjunctions (AND tests), all of which must pass in order for the rule to fire. In this sense, rules can be considered a generalization of decision trees and indeed a decision tree can trivially be summarized as a rule set (which tends to be a more compact and more intelligible representation of the tree) (Quinlan 1987, 1993). The opposite, however, does not hold: describing a rule as a tree is not immediate, for the tree imposes a more rigid logical structure than the rule.

The price paid for compactness is that several rules may trigger at the same time providing conflicting classifications; alternatively, no rule may fire at all. A number of techniques address these problems. Rules can be ordered and executed sequentially or majority voting can be used to resolve conflicting classifications. In the case where no rule fires, instances can be assigned to the most common class or to a special "unknown" class.

The performance of rules is quite similar to decision trees and even outperforms them on a variety of problems (Quinlan 1987; Pagallo and Haussler 1990; Weiss and Indurkya 1991; Cohen 1995). The advantage is that they provide a more compact and interpretable output that can be easily deployed in transaction monitoring systems.

A variety of rule discovery algorithms exist in the literature. One possible strategy is generating rule sets from decision trees. This is the approach followed in C4.5Rules (Quinlan 1993) and PART (Frank and Witten 1998). The former method generates rule sets from a large C4.5 decision tree. The latter builds a C4.5 tree in each iteration, making a rule out of the best leaf and then removing the matching observations and repeating the process. Sequential covering is an alternative strategy. One rule is learned at a time, all the matching observations are removed from the training dataset, and the process is repeated. RIPPER (Cohen 1995) generates a set of disjunctive rules for the minority class and then optimizes them by generating alternatives for each rule. Ripple-down rules (Gaines and Compton 1995) generate a rule on the majority class and then proceed adding exceptions.

More recently, the concept of ensemble learning was applied to rule algorithms also. RuleFit (Friedman and Popescu 2008) used rules generated from decision trees as base "weak learners" that are then combined in a weighted additive fashion instead of the more common disjunctive form.

4.2.4 Neural Networks. Neural networks were widely used in fraud detection, particularly in credit cards. They are the classifiers underlying commercial systems such as HNC's

Falcon System, acquired in 2002 by Fair Isaac's Corporation (Gopinathan et al. 1998) or Xtract's Detect, as well as in-house developed systems at financial institutions, such as Visa's CRIS System (Fryer 1996), Mellon Bank's FDS credit card fraud detection system (Ghosh and Reilly 1994), or Sociedad Espanola de Medios de Pago's (SEMP) Minerva system (Dorronsoro et al. 1997).

Typically, feed-forward networks with only three layers (input, hidden, and output layers) are used in fraud detection. The input to the neural network is the vector of features. The signal emitted by the output unit is the probability of the activity being criminal, which is used as a suspicion score. Given enough hidden units and proper nonlinearities and weights, three-layer neural nets are able to implement a universal function approximator (Haykin 1998). Backpropagation is commonly used for training (Haykin 1998). The weights are initialized with random values, which are then changed in the direction that minimizes training error. More complex setups with two hidden layers, or strategies other than backpropagation are possible, but uncommon. See Fadlalla and Lin (2001) for a survey of the techniques most commonly used in finance.

Neural networks are attractive in financial crime detection for a few reasons. First, three-layer nets were shown to be capable of dealing with the highly skewed class distributions that arise in this application. Dorronsoro et al. (1997) reported positive results of the Minerva system with ratios of fraud-to-legitimate transactions of 1 : 150. Second, once they are trained, they can analyze new data very quickly, an attribute that is necessary when trying to catch fraudulent transactions in real time.

However, neural networks also suffer from drawbacks. One major issue is the need to select and adjust the structure of the network. The choice of the number of hidden states must be made to optimize learning and generalization. Further, the performance of the classifier is very sensitive to the vector of features chosen, so significant attribute selection and preprocessing (e.g., normalization) are necessary. Maes, Tuyls, and Vanschoenswinkel (2002) reported a 28% improvement in true positive rate with a 10% false positive rate in a fraud detection experiment after removing just one correlated feature out of a set of 10 and normalization of the remaining features. In the money laundering example presented in Section 4.2.2, a three-layer network with four nodes in the hidden layer correctly classified 94.59% of the samples in the dataset. The false negative rate was 2.8% and the false positive rate was 10% when the dataset was log-normalized prior to training. If this preprocessing step is removed, the network converges to the trivial classifier that decides that all the instances are fraud free. Postprocessing can be a feasible option in some cases as well. Kim and Kim (2002) addressed the problem of overlapping data in credit card fraud by weighting the score of a neural network by a metric of fraud density in the neighborhood of the input feature vector.

Training neural networks is time consuming for large training datasets, especially if the model is intended to be retrained very often. In addition, backpropagation-trained multilayer perceptrons are prone to overfitting; a number of algorithms exist that address this problem, at the expense of adding complexity to the training process. Finally, neural networks are often treated as "black boxes," and their results can be difficult to interpret.

4.2.5 Bayesian Belief Networks. Bayesian belief networks (BBN) form another popular class of fraud detection methods. A BBN represents the joint probability distribution over a set of random variables as a directed acyclic graph; each node is a variable and arrows represent correlation between variables. A conditional probability table quantifies the effect of parent nodes on a child node. If a node has no parents, the table contains the prior probability of the variable. The probability of any configuration of system components can be calculated using the chain rule

$$P(s_n, s_{n-1}, s_{n-2}, \ldots, s_1) = \prod_{i=1}^{n} P(s_i | s_{i-1}, \ldots, s_1).$$

A transaction characterized by a vector of attributes X is classified as fraudulent (F) when $P(F|X) > P(\overline{F}|X)$, which, using Bayes' rule, can be reduced to comparing

$$P(x_n|x_{n-1}, \ldots, x_1, F) \cdots P(x_1|F)P(F) > P(x_n|x_{n-1}, \ldots, x_1, \overline{F}) \cdots P(x_1|\overline{F})P(\overline{F}).$$

Training a BBN involves first learning the structure of the network, then calculating the conditional probabilities for that structure from the data. This second step is trivial once the network structure is known. Structure is learned by either using a human expert or by exploring the space of potential networks. A variety of search strategies exist, including general-purpose search algorithms (e.g., simulated annealing, genetic algorithms) and BBN-specific search algorithms such as K2 (Cooper and Herskovits 1992) or tree augmented naïve Bayes (TAN) (Friedman, Geiger, and Goldszmidt 1997). For each network, the conditional probabilities are estimated and the network "goodness of fit" to training data is scored. The complexity of the network can be factored in the score using the Akaike information criterion (AIC) or metrics based on the minimum description length (MDL) criterion (Lam and Bacchus 1994) to avoid overfitting. Other metrics commonly used include the Bayesian–Dirichlet (BD) metric and specializations such as Bayesian–Dirichlet equivalent (BDe) (Heckerman, Geiger, and Chickering 1995). Alternative methods to score-based search include conditional independence tests (Verma and Pearl 1992) or assessing the mutual information between the input variables (Ezawa and Norton 1995). See Heckerman (1995) for a comprehensive tutorial on Bayesian belief networks.

Figure 5 shows a TAN BBN trained with the money laundering dataset introduced in Section 4.2.2, which correctly classifies 82.88% of the instances in the test dataset, with a 24% false positive rate and a 14% false negative rate. TAN works by first assuming conditional independence between all predictors—that is, starting with a naïve Bayes model and then considers adding one single dependency to each one of them.

The resulting network relates the transaction speed score (x_2) to the individual expected level of activity (x_3) and then x_3 to the peer comparison score in terms of transaction volume (x_1). Caution must be exercised when trying to draw causation conclusions from these results. First, there usually are multiple network structures with equivalent classification performance, and therefore, multiple explanations for the observed data. Different search algorithms with different parameters may converge to different networks. Second, each structure is learned based

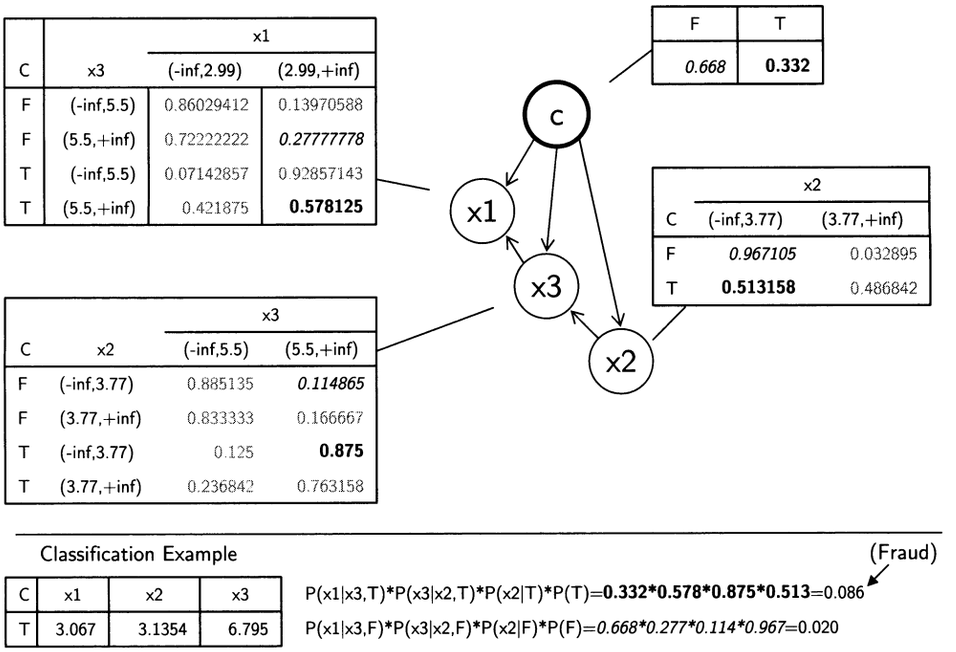

Figure 5. Bayesian belief network.

just on correlations observed in the training data; no causation can be taken for granted. Third, even if a causal relationship exists between two nodes of the net, there is still the problem of determining the actual direction of causation. In the example presented, according to expert judgement, it is indeed more plausible that a high transaction speed will result both in an unexpected level of activity and a large volume—even if the amount of each transaction is small.

Maes, Tuyls, and Vanschoenswinkel (2002) provided a comparative study between neural networks and Bayesian networks for credit card fraud detection. Transactions were described by four features and a label. BBN's yielded lower classification errors than neural networks, reporting that BBN's were faster to train but much slower to execute.

4.2.6 Hidden Markov Models. In a regular Markov model, the states follow a Markov process and are visible to the observer. In hidden Markov models (HMM's), the states are hidden. Instead, the observer views output variables that are influenced by the state and the sequence of output variables provide some information of the sequence of states.

As an illustration, we present an example to detect checking account overdraft fraud. Overdraft occurs when an account user conducts a withdrawal without sufficient funds in the account. Banks often provide a short-term lending service to their customers by allowing the withdrawal of funds even when the account does not have sufficient funds. For the financial institution, the downside of this service is the potential abuse that may occur in the form of nonpayment of the overdraft amount. Thus, to manage the risk of potential fraudulent activity while providing an overdraft service, a proper account monitoring approach must be employed. The customer's transaction activities can be modeled using a HMM. The following states may occur throughout the use of the account (shown in Figure 6):

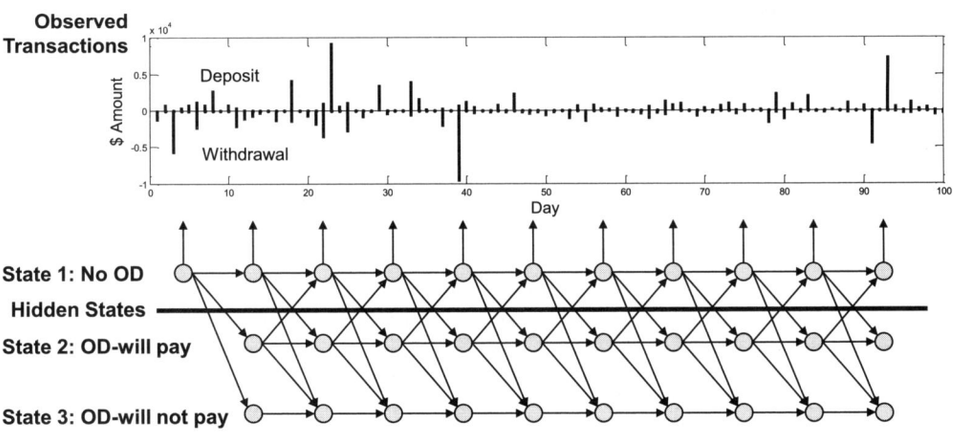

Figure 6. Transition states in HMM for overdraft classification.

- *State 1*: No overdraft; account has a nonnegative balance.
- *State 2*: Overdraft; account has a negative balance, but the customer has an intention to repay the overdraft at a later time.
- *State 3*: Overdraft; account has a negative balance and the customer has no intention to repay (i.e., fraudulent state).

A transition between the states is caused by a transaction made on the account. A transition from State 1 to itself happens when a customer, previously with a nonnegative balance, makes a transaction and remains in good standing. A customer can move from State 1 to either States 2 or 3 any time a withdrawal is made. If a customer is in State 2, and for some reason, the intention to repay the overdraft changes, a move to State 3 occurs. We think of State 3 as an "absorbing" state. That is, once a customer enters State 3, the customer remains there for all later times. These transitions from state to state form a Markov chain and the probability of transitions can be estimated from the historical time series of transaction data.

4.3 Link Analysis

Another major threat to financial institutions is organized crime rings, or groups of people working together to commit money laundering or fraud. The customers and their transactions, when viewed individually, may pass under the radar of normal detection schemes. This can happen either because they appear to be legal or because individual transactions involve small amounts of money. However, when the transactions are viewed in the context of a pattern of activity, often involving several related individuals, the criminal behavior can be more apparent.

One option for finding these groups is to use clustering algorithms to identify customers with similar behavioral patterns. However, these rings of criminal activity involve behaviors spread over many different transactions, over multiple customers and accounts, and often over lengthy periods of time. Ordinary cluster analysis may be unable to detect such complex networks. Link analysis and graph mining methods may be able to detect these groups of people working together. These techniques are common in areas including social sciences and law enforcement and they were recently applied in financial crime detection, especially for money laundering detection (Senator et al. 1995; Goldberg and Senator 1995, 1997; Goldberg and Wong 1998; Zhang, Salerno, and Yu 2003).

The main idea behind link analysis in this application is to start with a known entity of interest and find meaningful relationships with other entities. Often the attributes used to identify related individuals are defined by investigators based on their experiences. Zhang, Salerno, and Yu (2003) proposed a method for link discovery based on correlation analysis (LDCA), which they apply for investigating money laundering crimes. Here, a certain correlation is used to construct the attributes for link discovery methods.

Link analysis can also be used in "guilt-by-association" suspicion scoring (Macskassy and Provost 2005), where one entity gets a score that is a function of the scores of the entities it is associated to. This approach, however, may be very sensitive to different configuration parameters (Galstyan and Cohen 2005). See Macskassy and Provost (1997) for a comprehensive survey of the field and a case study of the implementation of a system for classification of networked data.

5. UNSUPERVISED LEARNING METHODS

One problem with supervised learning for financial crime detection is that the labels of class membership are often unreliable or unavailable. Previously worked cases are investigated by humans and can easily be mislabeled. In general, assigning labels to previous transactions and cases is time consuming and subject to errors. Further, in the case of money laundering, it is impossible to obtain class labels. There is no way to say with certainty that a customer has not committed money laundering. When there is no database of previous labeled cases available, unsupervised learning techniques must be used.

5.1 Clustering

Clustering algorithms segment the data into groups of similar observations, using some measure of similarity. While clustering on its own is not very useful for the detection of criminal behavior, standard clustering algorithms such as k-means were used in conjunction with supervised methods.

Clustering and profiling are combined in an unsupervised technique called peer group analysis (PGA) (Bolton and Hand 2001). In PGA, clusters of similar observations, called peer groups, are identified and their behaviors are characterized in the form of a profile. These peer groups are monitored over time and any member who starts to deviate from its peer group is noted and flagged for investigation.

5.2 Low-Dimensional Representation and Scoring

As mentioned before, one of the challenges in detecting and preventing financial crimes is the high-dimensionality of the data. Often a financial institution will like to know about groupings among observations and be able to identify observations that are most different from the others. However, data visualization in the full-dimensional space is infeasible and the analysis has to be done in the space that is useful for examining interesting behavior. Dimension reduction can help identify a small number of features that are most important for explaining patterns in the data. In addition, the score of the criterion function used for the transformation can also be used to provide a suspicion score for each observation.

When suspicion scoring is the goal, it is common to use either principal components analysis (PCA) or independent component analysis (ICA). Exploratory projection pursuit is another technique that is especially useful when the goal is to find clusters in the original data. Multidimensional scaling (MDS) is also used to determine the underlying dimensions that are useful for explaining similarities (or dissimilarities) between observations.

Figure 7 shows an analysis on a dataset of customer transaction histories using clustering and MDS. High-dimensional features are extracted to summarize transaction behaviors over time and the customers are first clustered based on these features. For the clustering, k-means clustering with $k = 12$ clusters is applied. The data are projected onto two-dimensional space using MDS. This approach clearly shows the unusual behavior of one cluster of customers.

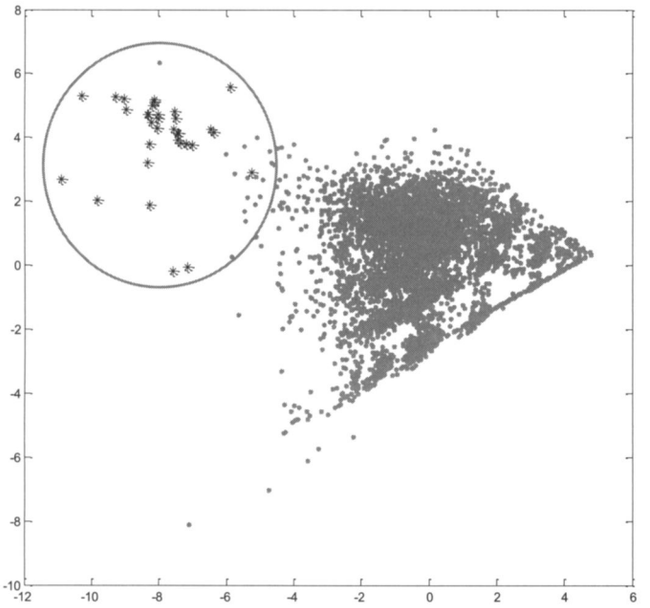

Figure 7. Clustering and multidimensional scaling projection of customer transactions to identify suspicious groups.

5.3 Anomaly Detection

In traditional statistical process control, anomalous observations may indicate that the process is out of control. In financial crimes, there are often observations (i.e., anomalies) in a dataset that does not belong to any cluster. In financial applications, anomalous observations often correspond to criminal transactions or customers and the reasons for their dissimilarities from the rest of the data can be useful for identifying the differences between criminal and noncriminal behavior. In the following discussion, we explore various anomaly detection methods.

Anomaly detection can be thought of as an application of outlier detection, which is a well-studied problem in the statistical community (Hawkins 1980; Barnett and Lewis 1994). This is a fairly straightforward idea for univariate data and there are standard statistical methods to find outliers in a dataset. In multivariate settings, there is no natural ordering or distance metric. The most common measure of distance that is used is the Mahalanobis distance, which was developed in the context of multivariate normal data. If one does not want to make strong parametric assumptions, the problem is more difficult.

5.3.1 Density and Distance-Based Outlier Detection. Several outlier detection algorithms were introduced in response to these needs and they fall into two main classes: density-based and distance-based. Density-based methods identify outliers as observations located in areas of low density. These methods find outliers as a by-product of clustering. A second class of algorithms uses distance-based methods. Here, outlying observations are identified as those that are "far" from some fraction of the remaining observations, according to some distance function. Finding distances between high-dimensional points for large datasets is computationally intensive, but there has been recent work on methods to do it efficiently (Knorr and Ng 1998). In both distance and density-based methods, it is necessary to define a distance metric.

Distance-based outlier detection was also extensively used in telecommunications fraud detection and these methods also generalize to financial crimes. Ferreira et al. (2006) characterized customer expected behavior by summarizing a number of feature variables (e.g., duration of calls and number of international calls) with one or two parameters per variable. An overall distance metric is computed as a linear combination of the individual distances observed on each of the feature variables. The Z-score and a Poisson-based score are used as distance metrics for features with two and one parameters, respectively. Murad and Pinkas (1999) proposed *CD-distance*, a metric of distance based on cumulative probability distributions to address limitations of distance metrics commonly used in pattern recognition (e.g., Euclidean, Hellinger, Mahalanobis, and divergence).

There was also work to develop a notion of data depth for nonparametric multivariate data analysis (Liu, Parelius, and Singh 1999). Observations are assigned a depth relative to the "center" of the dataset using a defined depth function. As it provides an ordering for data, it also leads to a class of multivariate outlier detection methods: the center of the data is the observation with maximum depth and outlying observations are those with minimum depth.

5.3.2 Detection of Anomalies in AML. The following example involves transaction features from a group of 100 customers. The goal is to identify customers with transactions that depart significantly from their peers. As an illustration, we use three features: x_1: average amount of cash transactions, x_2: structuring amount (amount of transactions below reporting limits), and x_3: ratio of structuring to other transactions. We use a depth-based detection technique by applying Mahalanobis distance (MD), modified to perform in a similar manner to the convex hull peeling strategy (Barnett and Lewis 1994).

In this approach, we apply MD to find the most extreme outlier, remove it, and repeat the process iteratively. This iterative peeling approach that removes one outlier at a time helps to overcome the well-known *masking* and *swamping* problems with MD. Masking is a situation where one outlier masks a second outlier. That is, only after the deletion of the first outlier will the second emerge as an outlier. This situation occurs when a cluster of outliers skews the mean and the covariance toward it and the resulting distance of the outliers from the mean becomes small. Swamping is a situation that occurs when the deletion of one outlier makes a second outlier look typical rather than unusual. This situation occurs when the presence of a cluster of outliers skews the mean and covariance away from other nonoutliers, making the nonoutliers look like outliers. Use of this iterative approach makes these problems less severe. The data points and the result of applying this strategy are shown in Figure 8. The sequence of the top 10 outliers found by the approach is labeled subsequently from $1, 2, \ldots, 10$ and highlighted by circles. This sequence of outliers becomes a useful prioritization tool to assign the order of priority for AML case investigation.

6. CONCLUSION

This article focuses on money laundering and retail banking fraud, two important types of financial crimes. It reviews some of the statistical challenges in these applications and discusses

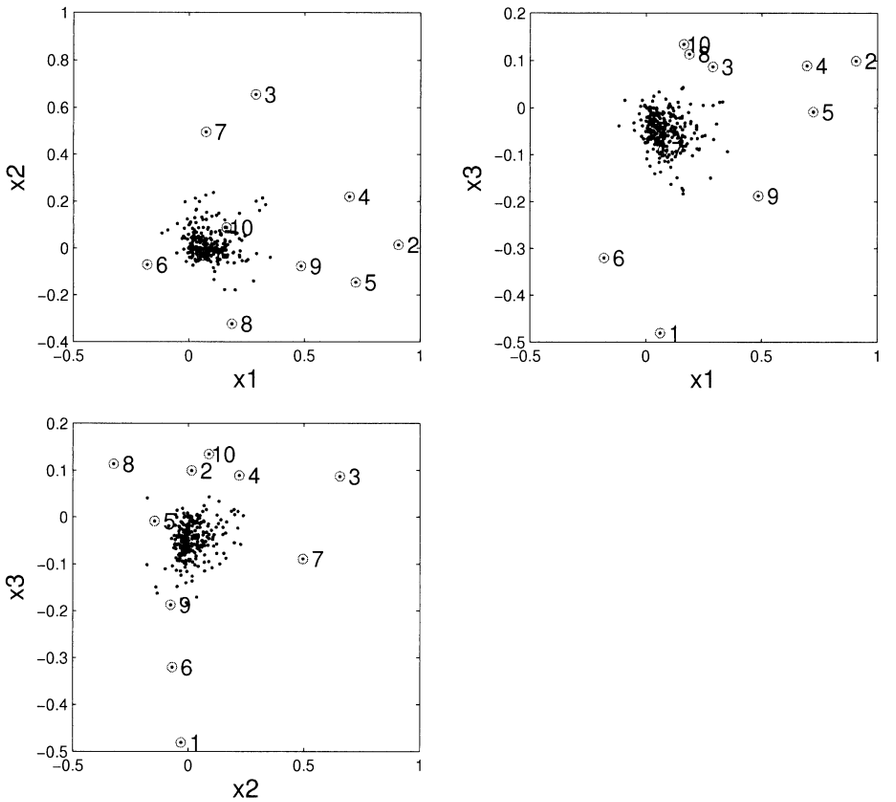

Figure 8. Outlier sequence identified by MD-peeling algorithm.

commonly used methods for fighting these financial crimes. While the existing methods were used effectively in some areas, further research and development are necessary to deal with the ever-changing nature of the problems.

We highlight methods for detecting financial crimes, but it is important to emphasize that prevention activities are equally, if not more important in the fight against these crimes. Prevention involves efforts by both customers and the institutions. Proper education can ensure that employees and customers are aware of steps needed to protect information that can lead to financial crime. Policies and procedures are needed in the industry to make sure that information is not misused in fraudulent activities and businesses must also take an active role in forming relationships with customers in industries where fraud is likely.

[Received March 2007. Revised April 2009.]

REFERENCES

Ahn, H., Lee, K., and Kim, K. (2006), "Global Optimization of Support Vector Machines Using Genetic Algorithms for Bankruptcy Prediction," in *Proceedings of ICONIP 2006, Part III. Lecture Notes in Computer Science*, Vol. 4234, New York: Springer, pp. 420–429. [10]

Aleskerov, E., Freisleben, B., and Rao, B. (1997), "CARDWATCH: A Neural Network-Based Database Mining System for Credit Card Fraud Detection," in *Proceedings of IEEE/IAFE on Computational Intelligence for Financial Engineering*, Amsterdam: Elsevier, pp. 220–226. [5]

Baena-García, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., and Morales-Bueno, R. (2006), "Early Drift Detection Method," in *ECML PKDD 2006 Workshop on Knowledge Discovery From Data Streams*, Berlin, Germany. [9]

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*, New York: Wiley. [16]

Bifet, A., and Kirkby, R. (2009), "Data Stream Mining: A Practical Approach," technical report, University of Waikato. [9]

Bolton, R. J., and Hand, D. J. (2001), "Unsupervised Profiling Methods for Fraud Detection," in *Credit Scoring and Credit Control VII*, Edinburgh, U.K. [15]

——— (2002), "Statistical Fraud Detection: A Review," *Statistical Science*, 17, 235–255. [6]

Breiman, L. (2001), "Random Forest," *Machine Learning*, 45, 5–32. [11]

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983), *CART: Classification and Regression Trees*, Belmont, CA: Wadsworth. [10]

Cahill, M. H., Lambert, D., Pinheiro, J. C., and Sun, D. X. (2002), "Detecting Fraud in the Real World," in *Handbook of Massive Data Sets*, Norwell, MA: Kluwer Academic, pp. 911–929, [10]

Caputo, B., Sim, K., Furesjo, F., and Smola, A. (2002), "Appearance-Based Object Recognition Using SVMs: Which Kernel Should I Use?" in *Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision*, Whistler, Canada. [10]

Carter, C., and Catlett, J. (1987), "Assessing Credit Card Applications Using Machine Learning," *IEEE Expert*, 2 (3), 71–79. [11]

Chan, P. K., Fan, W., and Prodromidis, A. L. (1999), "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems*, 14 (6), 67–74. [10,11]

Chawla, N. W., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321–357. [9]

Chen, R., Chiu, M., Huang, Y., and Chen, L. (2004), "Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines," in *Proceedings of IDEAL 2004. Lecture Notes in Computer Science*, Vol. 3177, New York: Springer, pp. 800–806. [10]

Chen, R., Luo, S., Liang, X., and Lee, V. (2005), "Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud," in *Proceedings of 2005 International Conference on Neural Networks and Brain*, Piscataway, NJ: IEEE, pp. 810–815. [10]

Cohen, W. (1995), "Fast Effective Rule Induction," in *Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA*, San Francisco, CA: Kaufmann, pp. 115–123. [12]

Cooper, G. F., and Herskovits, E. (1992), "A Bayesian Method for the Induction of Probabilistic Networks From Data," *Machine Learning*, 9, 309–347. [13]

Cortes, C., and Pregibon, D. (2001), "Signature-Based Methods for Data Streams," *Data Mining and Knowledge Discovery*, 5 (3), 167–182. [10]

Domingos, P. (1999), "MetaCost: A General Method for Making Classifiers Cost-Sensitive," in *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, New York: ACM. [9]

Dorronsoro, J., Ginel, F., Sanchez, C., and Santa Cruz, C. (1997), "Neural Fraud Detection in Credit Card Operations," *IEEE Transactions on Neural Networks*, 8 (4), 827–834. [9,13]

Drummond, C., and Holte, R. C. (2005), "Severe Class Imbalance: Why Better Algorithms Aren't the Answer," in *Proeedings of the 24th European Conference on Machine Learning*, Pisa, Italy, Porto, Portugal. [9]

Duda, R., Hart, P., and Stork, D. (2000), *Pattern Classification* (2nd ed.), New York: Wiley. [10]

Ezawa, K. J., and Norton, S. W. (1995), "Knowledge Discovery in Telecommunications Services Data Using Bayesian Models," in *Proceedings of the First International Conference of Knowledge Discovery & Data Mining*, pp. 100–105. [13]

Fadlalla, A., and Lin, C. H. (2001), "An Analysis of the Applications of Neural Networks in Finance," *Interfaces*, 31 (4), 112–122. [13]

Fawcett, T., and Provost, F. (1997), "Adaptive Fraud Detection," *Data Mining and Knowledge Discovery*, 1 (3), 291–316. [9,10]

Federal Bureau of Investigation, U.S. Department of Justice (2005), "Financial Crimes Report to the Public," public report, available at *http://www.fbi.gov/publications/financial/fcs_report052005/fcs_report052005.htm*. [5]

Ferreira, P. G., Alves, R., Belo, O., and Cortesao, L. (2006), "Establishing Fraud Detection Patterns Based on Signatures," in *Industrial Conference on Data Mining (ICDM)*, New York: Springer, pp. 526–538. [16]

Foster, D. P., and Stine, R. A. (2004), "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy," *Journal of the American Statistical Association*, 99, 303–313. [10]

Frank, E., and Witten, I. H. (1998), "Generating Accurate Rule Sets Without Global Optimization," in *Fifteenth International Conference on Machine Learning*, San Francisco, CA: Kaufmann, pp. 144–151. [12]

Freund, Y., and Schapire, R. (1997), "A Decision-Theoretic Generalization of Online Learning and Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139. [11]

Friedman, J. H., and Popescu, B. E. (2008), "Predictive Learning via Rule Ensembles," *Annals of Applied Statistics*, 2 (3), 915–954. [12]

Friedman, N., Geiger, D., and Goldszmidt, M. (1997), "Bayesian Network Classifiers," *Machine Learning*, 29 (2–3), 131–163. [13]

Fryer, B. (1996), "Visa Cracks Down on Fraud," *InformationWeek*, August 26, 1996. [13]

Gaines, B., and Compton, P. (1995), "Induction of Ripple-Down Rules Applied to Modeling Large Databases," *Journal of Intelligent Information Systems*, 5 (3), 211–228. [12]

Galstyan, A., and Cohen, P. R. (2005), "Is Guilt by Association a Bad Thing?" in *Proceedings of the Intelligence Analysis Conference*, McLean, VA. [15]

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004), "Learning With Drift Detection," in *Advances in Artificial Intelligence—SBIA 2004. Lecture Notes in Computer Science*, Vol. 3171, Berlin: Springer-Verlag, pp. 286–295. [9]

Gartner (2005), "Gartner Says ATM/Debit Card Fraud Resulted in $2.75 Billion in Losses in Past Year," Press Release, August 2, 2005; available online at *http://www.gartner.com/it/page.jsp?id=492168* (accessed: January 12, 2007). [5]

Ghosh, S., and Reilly, S. L. (1994), "Credit Card Fraud Detection With a Neural Network," in *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences*, Computer Society Press. [5,13]

Goldberg, H. G., and Senator, T. E. (1995), "Restructuring Databases for Knowledge Discovery by Consolidation and Link Analysis," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Menlo Park, CA: AAAI Press, pp. 136–141. [15]

——— (1997), "Break Detection Systems," in *AI Approaches to Fraud Detection and Risk Management: Collected Papers From the 1997 Workshop*, Technical Report WS-97-07, Menlo Park, CA: AAAI Press. [15]

Goldberg, H. G., and Wong, R. W. H. (1998), "Restructuring Transactional Data for Link Analysis in the FinCEN AI System," in *Papers From the 1998 Fail Symposium on Artificial Intelligence and Link Analysis, October 23–25, Orlando, FL*, Technical Report FS-98-0, Menlo Park, CA: AAAI Press, pp. 38–46. [15]

Gopinathan, K. M., Biafore, L. S., Ferguson, W. M., Lazarus, M. A., Pathia, A. K., and Jost, A. (1998), "Fraud Detection Using Predictive Modeling," U.S. Patent 5819226, October 6. [13]

Hawkins, D. (1980), *Identification of Outliers*, London: Chapman & Hall. [16]

Haykin, S. (1998), *Neural Networks: A Comprehensive Foundation* (2nd ed.), Englewood Cliffs, NJ: Prentice Hall. [13]

Heckerman, D. (1995), "A Tutorial on Learning Bayesian Networks," Technical Report MSR-TR-95-06, Microsoft Research. [13]

Heckerman, D., Geiger, D., and Chickering, D. M. (1995), "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, 20 (3), 197–243. [13]

Hothorn, T., Hornik, K., and Zeileis, A. (2006), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15 (3), 651–674. [11]

Hsu, C., Chang C., and Lin, C. (2009), "A Practical Guide to Support Vector Classification," technical report, National Taiwan University, Dept. of Computer Science. [10]

Kass, G. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29 (2), 119–127. [10]

Kim, M. J., and Kim, T. S. (2002), "Neural Classifier With Fraud Density Map for Effective Credit Card Fraud Detection," in *IDEAL 2002. Lecture Notes in Computer Science*, Vol. 2412, New York: Springer, pp. 378–383. [13]

Klinkenberg, R., and Uping, S. (2002), "Concept Drift and the Importance of Examples," in *Text Mining: Theoretical Aspects and Applications*, eds. J. Franke, G. Nakhaeizadeh, and I. Renz, Berlin: Springer. [9]

Knorr, E. M., and Ng, R. T. (1998), "Algorithms for Mining Distance-Based Outliers in Large Datasets," in *Proeedings of the 24th VLDB Conference*, New York, USA. [16]

Kohavi, R. (1997), "Wrappers for Feature Subset Selection," *Artificial Intelligence*, 97, 273–324. [10]

Lam, W., and Bacchus, F. (1994), "Learning Bayesian Belief Networks. An Approach Based on the MDL Principle," *Computational Intelligence*, 10, 269–293. [13]

Lee, T., and Chen, I. (2003), "Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines," in *International Conference on Information and Knowledge Engineering—IKE '03*, Vol. 2, Las Vegas, NV: CSREA Press, pt. 2, pp. 533–538. [11]

Li, F., Xu, J., Dou, Z., and Huang, Y. (2004), "Data Mining-Based Credit Evaluation for Users of Credit Card," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, Piscataway, NJ: IEEE. [11]

Liu, R., Parelius, J., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics, and Inference," *The Annals of Statistics*, 27, 783–858. [16]

Macskassy, S. A., and Provost, F. (1997), "A Brief Survey of Machine Learning Methods for Classification in Networked Data and an Application to Suspicion Scoring," in *Lecture Notes in Computer Science*, Vol. 4503, Berlin: Springer-Verlag, pp. 172–175. [15]

——— (2005), "Suspicion Scoring Based on Guilt-by-Association, Collective Inference, and Focused Data Access," in *Proceedings of the International Conference on Intelligence Analysis*, McLean, VA. [15]

Maes, S., Tuyls, K., and Vanschoenswinkel, B. (2002), "Credit Card Fraud Detection Using Bayesian and Neural Networks," in *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies*, Havana, Cuba. [13,14]

Mercer, L. C. J. (1990), "Fraud Detection via Regression Analysis," *Computers and Security*, 9 (4), 331–338. [10]

Montana, G. (2006), "Balancing Risk and Reward in the Deposits Business," *The RMA Journal*, October 2006. [6]

Murad, U., and Pinkas, G. (1999), "Unsupervised Profiling for Identifying Superimposed Fraud," in *PKDD'99: Principles of Data Mining and Knowledge Discovery*, Berlin: Springer-Verlag, pp. 15–18. [16]

Opitz, D., and Maclin, R. (1999), "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, 11, 169–198. [12]

Pagallo, G., and Haussler, D. (1990), "Boolean Feature Discovery in Empirical Learning," *Machine Learning*, 5 (1), 71–99. [12]

Quinlan, J. R. (1987), "Simplifying Decision Trees," *International Journal of Man–Machine Studies*, 27 (3), 221–234. [12]

——— (1993), *C4.5: Programs for Machine Learning*, San Fransisco: Kaufmann. [10,12]

Rogers, J., and Gunn, S. (2005), "Identifying Feature Relevance Using a Random Forest," in *Proceedings of Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimisation Perspectives Workshop*, Bohinj, Slovenia. [12]

Senator, T. E., Goldberg, H. G., Wooton, J., Cottini, M. A., Umar Khan, A. F., Klinger, C. D., Llamas, W. M., Marrone, M. P., and Wong, R. W. H. (1995), "The FinCEN Artificial Intelligence System (FAIS): Identifying Potential Money Laundering From Reports of Large Cash Transactions," *AI Magazine*, 16 (4), 21–39. [15]

Siddiqi, N. (2005), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, New York: Wiley. [10]

Street, W. N., and Kim, Y. (2001), "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, pp. 377–382. [9]

Tang, J., and Yin, J. (2005), "Developing an Intelligent Data Discriminating System of Anti-Money Laundering Based on SVM," in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Piscataway, NJ: IEEE. [10]

Vapnik, V. N. (2000), *The Nature of Statistical Learning Theory* (2nd ed.), Berlin: Springer-Verlag. [10]

Verma, T., and Pearl, J. (1992), "An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation," in *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Kaufmann, pp. 323–330. [13]

Wang, H., Fan, W., Yu, P. S., and Han, J. (2003), "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," in *ACM SIGKDD*, August 2003, New York: ACM. [9,10]

Weiss, S. M., and Indurkhya, N. (1991), "Reduced Complexity Rule Induction," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, San Mateo, CA: Kaufmann. [12]

Widmer, G., and Kubat, M. (1996), "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning*, 23, 69–101. [9]

Witten, I. A., and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.), San Francisco: Kaufmann. [10]

Zhang, Z., Salerno, J. J., and Yu, P. S. (2003), "Applying Data Mining in Investigating Money Laundering Crimes," in *Proceedings of the Ninth ACM SIGKDD*, New York: ACM, pp. 747–752. [15]