7-2010

Detecting Management Fraud in Public Companies

Mark Cecchini
University of South Carolina - Columbia, Cecchini@moore.sc.edu

Haldun Aytug

Gary J. Koehler

Praveen Pathak

Publication Info

Published in *Management Science*, Volume 56, Issue 7, 2010, pages 1146-1160.
http://www.informs.org/Find-Research-Publications/Journals

MANAGEMENT SCIENCE
Vol. 56, No. 7, July 2010, pp. 1146–1160
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5607 | 1146

informs

DOI 10.1287/mnsc.1100.1174

Detecting Management Fraud in Public Companies

Mark Cecchini
School of Accounting, Darla Moore School of Business, University of South Carolina, Columbia, South Carolina 29208, cecchini@darla.moore.sc.edu

Haldun Aytug, Gary J. Koehler, Praveen Pathak
Information Systems and Operations Management, Warrington College of Business Administration, University of Florida, Gainesville, Florida 32611 {aytugh@ufl.edu, koehler@ufl.edu, praveen@ufl.edu}

This paper provides a methodology for detecting management fraud using basic financial data. The methodology is based on support vector machines. An important aspect therein is a kernel that increases the power of the learning machine by allowing an implicit and generally nonlinear mapping of points, usually into a higher dimensional feature space. A kernel specific to the domain of finance is developed. This financial kernel constructs features shown in prior research to be helpful in detecting management fraud. A large empirical data set was collected, which included quantitative financial attributes for fraudulent and nonfraudulent public companies. Support vector machines using the financial kernel correctly labeled 80% of the fraudulent cases and 90.6% of the nonfraudulent cases on a holdout set. Furthermore, we replicate other leading fraud research studies using our data and find that our method has the highest accuracy on fraudulent cases and competitive accuracy on nonfraudulent cases. The results validate the financial kernel together with support vector machines as a useful method for discriminating between fraudulent and nonfraudulent companies using only publicly available quantitative financial attributes. The results also show that the methodology has predictive value because, using only historical data, it was able to distinguish fraudulent from nonfraudulent companies in subsequent years.

Key words: management fraud; classification; support vector machines; financial event detection; kernel methods
History: Received December 19, 2005; accepted February 25, 2010, by Stefan Reichelstein, accounting. Published online in *Articles in Advance* May 17, 2010.

1. Introduction and Motivation

Statement on Auditing Standards (SAS) 99, "Consideration of Fraud in a Financial Statement Audit" (AICPA 2002), establishes external auditors' responsibility to plan and perform audits to provide a reasonable assurance that the audited financial statements are free of material fraud. Recent events highlight that failing to detect fraudulent financial reporting not only exposes the audit firm to adverse legal consequences, but also exposes the audit profession to increased public and governmental scrutiny. This has led to fundamental changes in the structure of the public accounting industry and government oversight of the accounting profession (e.g., consider the *Sarbanes-Oxley Act of 2002*, the creation of the Public Company Accounting Oversight Board in 2002, and subsequent actions of the New York Stock Exchange (2003)). Research that helps auditors better assess the risk of material misstatement during their planning phase may help reduce instances of fraudulent reporting. Such research is of interest to academicians, standard setters, regulators, audit firms, and investors.

Current research in accounting has examined methods to assess the risk of fraudulent financial reporting. The methodologies are varied and sometimes combine behavioral and quantitative factors. For example, Loebbecke et al. (1989) compiled an extensive list of company characteristics associated with fraudulent reporting ("red flags"). Hansen et al. (1996) and Bell and Carcello (2000) utilized red flag data to develop qualitative response and logistic regression models, respectively. These studies rely on information that can only be gathered via close personal contact, such as interviews. Other methods rely only on publicly-available data, including quantitative financial variables and indicator variables, such as auditor tenure (Green and Choi 1997, Summers and Sweeney 1998, Beneish 1999, Dechow et al. 2009).

This paper proposes a methodology to aid in detecting fraudulent financial reporting by utilizing only basic and publicly available financial data. Our approach combines aspects of the fraud assessment research in accounting with computational methods and theory used in machine learning/datamining. We gather a sample of 205 fraudulent companies using accounting and auditing enforcement releases (AAERs). We match our fraud sample with 6,427

nonfraudulent companies. Using these data, we validate our approach. We also replicate the results of other leading fraud detection studies using our data and compare the outcomes. Furthermore, we discuss which variables were important in the resulting output to gather new insights for auditors from our proposed methodology. The resulting decision aid has the potential to complement the unaided auditor risk assessments envisioned in SAS 99.

Machine learning uses computational techniques to automate the discovery of patterns that may be difficult to find otherwise. Machine learning methodologies have been used to determine financial statement validity and the likelihood of bankruptcy and credit worthiness (for example, Green and Choi 1997, Tam and Kiang 1992). There are many models commonly used in machine learning such as neural networks (NNs) (Haykin 1998), linear discriminant functions (Fisher 1936), logit functions (Agresti 1990), and decision trees (Quinlan 1996). Attempts have been made to recognize patterns in fraudulent companies using neural networks (Fanning et al. 1995, Green and Choi 1997), probit functions (Beneish 1999), logit functions (Summers and Sweeney 1998, Bell and Carcello 2000, Dechow et al. 2009), and expert systems (Ragothaman et al. 1995). These studies utilized publicly available quantitative data from financial statements and, in some cases, surveys from auditors.

What distinguishes our proposed approach from prior attempts to understand and aid fraud-risk assessments is our use of recent advances in machine learning theory, both through statistical learning theory that addresses generalization errors (Vapnik 1995) and by utilizing methods that facilitate incorporating domain knowledge (with a nonlinear mapping called a kernel) while the learning task is undertaken. The kernel we construct (which we call the financial kernel (FK)) is crafted to map observable financial attributes to nonlinear features shown effective in prior research. We then use our financial kernel with a powerful machine learning technique, called support vector machines (SVMs), that implements statistical learning theory. The methodology we create can be generalized to other accounting issues, such as the early detection of bankruptcy, prediction of restatements, early detection of increased market value, and general industry stability. For this study, we focus on the management fraud problem.

In the following section we review fraud detection literature and summarize key concepts and results. Section 3 briefly reviews support vector machines and statistical learning theory. Section 4 builds on §3, explaining the concept of a kernel, the mechanism used to incorporate domain-specific knowledge in support vector machines. In §5, we develop our financial kernel. In §6, we explain the data gathering process, the testing methodology, and prediction results of the FK with SVM on an empirical data set that includes data on fraudulent and nonfraudulent companies (we will often abbreviate the term fraudulent with "fraud" and nonfraudulent with "nonfraud"). An analysis of the results is given as well. Section 7 concludes this paper and gives directions for future research.

2. Fraud Detection Literature

A common thread in previous related literature is an attempt to find indicators of potential fraud, sometimes called red flags. This literature is generally targeted at the auditing profession because most red flag studies (Loebbecke et al. 1989, Pincus 1989, Asare and Wright 2004) focus on information that can only be determined at close contact. A key result was given by Loebbecke et al. (1989). They partitioned a large set of indicators into three main components: conditions, motivation, and attitude. They found in 86% of the fraud cases that at least one factor from each component was present, indicating that it is extremely rare for fraud to exist without all three components existing simultaneously. Hackenbrack (1993) found that the relative influence of such red flags on auditor fraud-risk assessments varies systematically with auditor experience. This subjectivity affects the red flag decision and ultimately the assessed risk of fraud. Pincus (1989) and Asare and Wright (2004) find that auditors using a standardized red flag program are less successful at correctly identifying fraud risk. These experimental studies show that there are some problems with the red flag checklist as a fraud detection mechanism.

Bell and Carcello (2000) developed a logistic regression model to estimate the likelihood of fraudulent financial reporting using the red flag data from the Loebbecke et al. (1989) study. The model scored higher than auditing professionals using their own judgment for the detection of fraud and was able to achieve 81% accuracy predicting fraud. The effectiveness of this method is limited because inside information is needed to create the red flag checklist and subjective evaluations of the resulting information are needed.

Hansen et al. (1996) developed a generalized qualitative-response model to analyze management fraud. Over 20 trials they achieved an 89.3% predictive accuracy when the costs were assumed to be symmetric between fraudulent and nonfraudulent companies. After adjusting the model to allow for asymmetric costs, the overall accuracy dropped to 85.5%; however, the per-class error for fraudulent companies decreased markedly. The accuracy of this method validates the qualitative response model as a mechanism

for detecting fraud. As in Bell and Carcello (2000) and Loebbecke et al. (1989), the red flag checklist of inside information is the source of the variables. This research elucidated the importance of per-class accuracy in fraud detection research. Minimizing the error on the fraudulent class makes the model less likely to miss an actual fraudulent company, possibly at the expense of falsely labeling a nonfraudulent company as fraudulent. Given the high cost of missing a true fraudulent company, asymmetric costs may be preferable.

Algorithmic approaches tend to focus on the detection of fraud based on a mathematical function that includes attributes determined to be salient. Because there are no formal theoretical indicators of fraud (Green and Choi 1997), the attributes are usually chosen with some expert judgment, but on an ad hoc basis. Green and Choi (1997) used NNs to identify fraud using publicly available data. They trained the network on five ratios that were previously identified in auditor risk assessments for the revenue collection cycle. The model predicted fraud with 74.03% accuracy on a holdout sample. Although the results showed promise, the sample size was small (46 fraud, 49 nonfraud). Also, the neural network has a major limitation as a black box approach because examining the function that determines the cutoff between fraudulent and nonfraudulent cases is nontrivial.

Summers and Sweeney (1998) developed a cascaded logit model that considers financial variables together with variables that indicate insider trading of company shares. The insider trading cues are used to help discriminate fraudulent from nonfraudulent firms. The authors developed a matched sample of 51 fraudulent firms and 51 nonfraudulent firms, and achieved 72% accuracy in predicting fraud cases. The main weakness was the lack of a holdout sample. The results given are for the training set only. It is impossible to predict how well this model would generalize without some out-of-sample validation.

Another thread in the literature is to use statistical methods, such as probit or logistic regression, to predict fraud using publicly available data in a sample that closely mimics the relative sizes of the two classes of companies. Beneish (1999) developed a probit model and a weighted exogenous sampling maximum likelihood (WESML) model using eight quantitative financial variables to predict fraud. Five of the eight variables involved year-over-year changes (an important component of the financial kernel explained in §5). The study attempted to approximate the relative sizes of the two classes of companies with a skewed data set consisting of 50 fraudulent companies and 1,758 nonfraudulent companies. The probit model achieved 76% accuracy on manipulators in the estimation sample and 56.1% accuracy in the holdout sample. Beneish (1999) developed a method of model validation, showing how the probit and WESML models would compare to a naïve strategy. We follow this method of validation in §6.3.

Dechow et al. (2009) created an extensive database of fraud firms (680) based on AAERs spanning from 1982 to 2005. This study examines the characteristics of firms that manipulate financial results and compares them to nonmanipulating firms. Several dimensions of publicly available data are considered, including financial variables, market-related variables, and off-balance sheet and other nonfinancial variables. The characteristics that tend to best distinguish between fraud and nonfraud firms are accrual quality measures and firm performance measures, operating leases, abnormal changes in employees, order backlog, prior stock price performance, the earnings-to-price ratio, and the amount of new financing. The variables were then used in a logit model to test prediction accuracy. The highest accuracy achieved on a sample of 29,159 firms (133 manipulators) was 71.53% overall. The best Type I error[1] for the prediction models of Dechow et al. (2009) was 35.48%. We take the variables that are used in Dechow et al. (2009) along with variables from Beneish (1999), Summers and Sweeney (1998), and Green and Choi (1997) and apply our methodology with them. The goal of our research is to create the best overall prediction while at the same time controlling for Type I error.

Methods of determining generalization ability are lacking in some fraud detection research (e.g., in Summers and Sweeney 1998). It is one thing to show that a model is accurate when tested on the same set it is trained on, but this is seldom indicative of performance on out-of-sample cases. Tsai and Koehler (1993) tested the robustness of the results of several papers that used inductive learning and warned that the true accuracy of concepts learned by induction may not be revealed in studies of small sample size. There is a trade-off between the number of variables that are used for a model and the model's ability to correctly predict in the future. This concern is one that is often leveled against induction methods. In the next section we provide an overview of statistical learning theory and support vector machines and show how this methodology deals with this generalization issue.

3. Statistical Learning Theory and Support Vector Machines

Most machine learning/datamining methods use a training set of data having known positive and

[1] The Type I error as defined in this paper is (*number of fraud firms classified as nonfraud firms*)/(*total fraud firms*). The Type II error is defined as (*number of nonfraud firms classified as fraud firms*)/(*total nonfraud firms*). This definition was also used in Beneish (1999).

negative examples of the concept to be learned. This is called supervised learning. For example, if we are trying to learn how to discriminate between companies likely to default on loans from those unlikely to default, we would collect past cases of defaulting and nondefaulting companies as done in studies such as Abbot et al. (2004) and Messier and Hansen (1988). Such a training set consists of ℓ observations and a classification for each; that is, the training set $S = \{(\mathbf{u}^1, y^1), \ldots, (\mathbf{u}^\ell, y^\ell)\}$, where $\mathbf{u}^i \in X \subseteq \Re^n$ represent the ith observation with each having n attributes (the independent variables), X is the instance space of all possible companies, and $y^i \in \{-1, +1\}$ is the label representing negative and positive examples of the concept, respectively. (For the concept "companies likely to default on loans," a label of $+1$ means the company instance is a company likely to default on loans.) Unless otherwise stated, a vector is denoted by a bold, lowercase letter. The superscript on the vector denotes the instance. An unbolded, subscripted, lowercase letter refers to the components of the vector. The subscript represents the index of the component. The inner product of two vectors, $\mathbf{x}$ and $\mathbf{y}$, is represented by $\langle \mathbf{x}, \mathbf{y} \rangle$. In §5 we add a second subscript for the year (or period). Finally, $\mathbf{x}'$ is the transpose of vector $\mathbf{x}$.

Typical supervised learning approaches using, for example, neural networks or logit, start with a training set and best fit the training data given the concept structure chosen (a neural network or a logit function, respectively). This goal is known as empirical risk minimization—it focuses solely on reducing the error over the training set. As is well known, empirical risk minimization often results in overfitting (Eisenbeis 1987); that is, for small sample sizes, a small empirical risk does not guarantee a small overall risk of error over all possible cases in the instance space X. To ameliorate this, the data set is often broken into two (or more) sets where part of the cases are used for training to fit a function and the remaining part, the holdout, to test its ability to predict on a data set not used for fitting. These approaches help detect overfitting but do not eliminate it.

3.1. Statistical Learning Theory

Statistical learning theory (Vapnik 1995) addresses the overfitting problem formally and directly. Vapnik shows that learning a function from examples can be formulated as minimizing the overall risk functional. This risk functional is the expected loss over *all* the instance space when using a given induced function. The expected loss involves an unknown sampling distribution. Vapnik (1995) handled this by first developing a bound on the expectation that is tight for some distributions, but valid for all. Second, he proposed the structural risk minimization principle for induction that focuses on minimizing this bound. Minimizing this bound has the practical effect of trading off empirical risk with generalization ability.

Support vector machines are a methodology that employs this process. In the next section we briefly touch on the more salient features of SVM methodology. A nice coverage can be found in Cristianini and Shawe-Taylor (2000).

3.2. Support Vector Machines

Support vector machines determine a hyperplane $\{\mathbf{x}: \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ in the feature space that best separates positive from negative examples. A feature space results from mapping the observable attributes to properties (i.e., features) that might better relate to the problem at hand. For example, given attributes price and earnings, the price-to-earnings (PE) ratio feature is easily constructed. For cases where the concepts are not linear, attributes can be easily mapped to nonlinear features using a kernel (we discuss this in more detail in §4).

Vapnik (1995) has shown that his structural risk minimization principle can be attained by SVMs. The following is the primal formulation (without a kernel) for a two-norm version of SVMs:

$$\min_{\xi^i \geq 0} \tfrac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + \sum_{i=1}^{\ell} C_{y_i} \xi^i \quad \text{s.t.} \quad y^i(\langle \mathbf{w}, \mathbf{u}^i \rangle + b) \geq 1 - \xi^i, \quad i = 1, \ldots, \ell.$$

The first term of the objective gives a measure of the complexity of the classification function, whereas the second term measures the empirical error over the training set. Vapnik's (1995) bound is a sum of these two terms. Here ξ^i is the ith slack variable, which allows for a classification error for the ith sample, thus allowing misclassifications on the training set; $\langle \mathbf{w}, \mathbf{w} \rangle$ is the squared two-norm of $\mathbf{w}$; parameters C_{+1} and C_{-1} are trade-offs between empirical errors ξ^i and generalization $\langle \mathbf{w}, \mathbf{w} \rangle$ appearing in Vapnick's bound; and the constraints try to put positive cases at a positive distance from the hyperplane and negative cases on the other side. When the data are linearly separable, the inclusion of the ξ slack variables is unnecessary. Because this two-norm problem is a convex quadratic program, SVM learning is theoretically guaranteed to find an optimal solution. Neural networks, decision trees, etc., do not carry this guarantee and have local optima leading to a plethora of heuristic approaches to find acceptable results. SVMs also scale up to very large data sets (Yu et al. 2003) and have been applied to problems involving text data (Joachims 1998), pictures (Shawe-Taylor and Cristianini 2004), etc.

This problem has a useful dual formulation. The dual formulation is

$$\max_{0\le\lambda^i\le C_{y^i}} \sum_{i=1}^{\ell}\lambda^i - \tfrac{1}{2}\sum_{i,j=1}^{\ell} y^i y^j \lambda^i \lambda^j \langle \mathbf{u}^i, \mathbf{u}^j\rangle$$
$$\text{s.t. } \sum_{i=1}^{\ell} y^i \lambda^i = 0, \tag{1}$$

where λ^i are the dual variables (Cristianini and Shawe-Taylor 2000). The dual solution is useful because the explicit usage of the data points is collapsed into a matrix of inner products, thus hiding the size of these vectors. When nonlinear mappings to features using kernels are employed, it is not uncommon for there to be an infinite number of features. The dual "hides" these infinities. The dual also allows for a generalization to kernel mappings, as we will discuss.

4. Kernel Methods

A kernel is an implicit mapping ϕ of an input attribute space X onto a (usually higher dimensional) feature space F. The kernel often improves the computational power of the learning machine by implicitly allowing combinations and functions of the original input attributes forming features, hence creating a nonlinear decision surface. Kernels provide a mechanism that helps to unravel spaces not linearly separable to ones that are potentially linearly separable. For example, if only price and earnings are inputs, a PE ratio would not be explicitly considered by a linear learning mechanism. A kernel, properly chosen, would allow many different relationships between variables to be simultaneously examined, presumably including price divided by earnings. The PE measure is a "feature" of the input variables. There are many powerful, generic kernels (Cristianini and Shawe-Taylor 2000, Genton 2001), but kernels can also be constructed to represent domain knowledge about a specific application area. Research suggests that kernels that are constructed with the help of application-specific information tend to have better results (Cristianini and Shawe-Taylor 2000). In line with this, we develop a financial kernel that allows for features usually observed in finance-related learning tasks. One type of feature incorporates ratios of input variables. Another common type of financial feature includes changes in ratios over time.

The kernel matrix is a matrix with entries $K_{ij} = \langle \phi(\mathbf{u}^i), \phi(\mathbf{u}^j)\rangle$, where ϕ is a mapping $\phi\colon X \to \Re^m$, and $\mathbf{u}^i, \mathbf{u}^j \in X$. Often the dimension of the feature space, m, is much larger than the attribute space, and may even be infinite. The objective of the dual formulation expressed in the objective of Equation (1) can be generalized to allow the usage of kernels as follows:

$$\max \sum_{i,j=1}^{\ell} y^i y^j \lambda^i \lambda^j K(\mathbf{u}^i, \mathbf{u}^j). \tag{2}$$

The kernel function is an inner product between feature vectors and is denoted as $K(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v})\rangle$, where $\mathbf{u}, \mathbf{v} \in X$. The feature vectors do not need to be explicitly calculated because the kernel function creates a mapping implicitly. The SVM formulation requires that a kernel function be symmetric (i.e., $K(\mathbf{u}, \mathbf{v}) = K(\mathbf{v}, \mathbf{u})$), be positive semidefinite, and satisfy the Cauchy–Schwarz inequality (Cristianini and Shawe-Taylor 2000).

There are many generic kernels in existence, and the list is ever growing (Cristianini et al. 2002, Joachims 1998, Rüping 2001). The polynomial kernel can be used to illustrate the nature and expressive power of these functions. The polynomial kernel is defined as $\hat{K}(\mathbf{u}, \mathbf{v}) = (K(\mathbf{u}, \mathbf{v}) + R)^d$, where $K(\mathbf{u}, \mathbf{v})$ is the normal inner product $\langle \mathbf{u}, \mathbf{v}\rangle$ kernel, d is a positive integer, and R is fixed. Consider two observations $\mathbf{u} = (u_1, u_2, u_3, u_4)'$ and $\mathbf{v} = (v_1, v_2, v_3, v_4)'$ with $d = 1$ and $R = 0$. Then $K(\mathbf{u}, \mathbf{v}) = u_1v_1 + u_2v_2 + u_3v_3 + u_4v_4$ and, with $R = 0$ and $d = 2$, $\hat{K}(\mathbf{u}, \mathbf{v}) = (u_1v_1 + u_2v_2 + u_3v_3 + u_4v_4)^2$. The kernel $K(\mathbf{u}, \mathbf{v})$ has four features and $\hat{K}(\mathbf{u}, \mathbf{v})$ has 10 features, u_1^2, u_2^2, u_3^2, u_4^2, u_1u_2, u_1u_3, u_1u_4, u_2u_3, u_2u_4, and u_3u_4. An alternative to using a kernel is to explicitly create all features as direct input to the SVM. For example, to emulate $\hat{K}(\mathbf{u}, \mathbf{v})$, one would compute all 10 features for every observation and use these as input.

A compelling property of kernel methods is the ability to form new kernels from existing kernels. Cristianini and Shawe-Taylor (2000) show that the set of kernel functions are closed under addition, multiplication, and scaling by a positive constant. In the next section we use such operators to create our financial kernel (while relegating a majority of the technical details to the appendix).

5. The Financial Kernel

Defining a domain-specific kernel for financial applications entails looking to the finance and accounting literature to see which attributes and features are often utilized for classification. Most such analyses look at ratios of items on the financial statements. Models for earnings quality in accounting utilize ratios (e.g., Francis et al. 2005). Loebbecke et al. (1989) used financial ratios as part of their management fraud model as well. All of the studies detailed in §2 used financial ratios.

Also, changes in ratios over time are important features found in this domain. McNichols and Wilson (1988) used year-over-year changes in key account values to help determine earnings management. Francis et al. (2005) utilized year-over-year changes extensively in their study on earnings quality. Beneish

Table 1 Two Hypothetical Firms

	A_1	A_2	L_1	L_2
Firm 1	10	11	12	13
Firm 2	4	6	8	10

(1999) utilized year-over-year changes to help determine management fraud. Year-over-year changes in ratios are captured by the function

$$\frac{(u_{i2}/u_{j2} - u_{i1}/u_{j1})}{u_{i2}/u_{j2}} = 1 - \frac{u_{i1}u_{j2}}{u_{j1}u_{i2}},$$

where $i, j = 1, \ldots, n$ are the attribute numbers with $i < j$. The second subscript represents the year (1 or 2). Note, the feature of relevance is $u_{i1}u_{j2}/u_{j1}u_{i2}$ because the constants and signs will be collapsed into the intercept of the linear discriminant function that SVM will induce. We denote the financial kernel as $K_F(\mathbf{u}, \mathbf{v})$ and give its detailed development in the appendix. Briefly, it computes all ratios of input attributes as well as year-over-year ratios.

The FK produces $3n(n-1)$ features starting with n attributes. The features can be broken down into six feature types. These six are shown below and represent intrayear ratios as well as year-over-year changes in ratios:

$$\phi(\mathbf{u}) = \left(\frac{u_{i1}}{u_{j1}}, \frac{u_{j1}}{u_{i1}}, \frac{u_{j2}}{u_{i2}}, \frac{u_{i2}}{u_{j2}}, \frac{u_{i1}u_{j2}}{u_{j1}u_{i2}}, \frac{u_{j1}u_{i2}}{u_{i1}u_{j2}}\right)', \quad i, j = 1, \ldots, n,\ i < j.$$

A simple example illustrates this. Assume we have two financial attributes, accounts receivable (A) and current liabilities (L), observed over two years. We can represent an observation using two attributes over two years, A_1, A_2, L_1, and L_2, respectively. Assume now that we have the following data for two firms.

Attribute vectors are $\mathbf{u}' = (10, 11, 12, 13)$ and $\mathbf{v}' = (4, 6, 8, 10)$ corresponding to each firm. Our kernel implicitly computes the following mapping for each firm:

$$\phi(\mathbf{u}) = \left(\frac{a_1}{l_1}, \frac{l_1}{a_1}, \frac{l_2}{a_2}, \frac{a_2}{l_2}, \frac{a_1 l_2}{a_2 l_1}, \frac{l_1 a_2}{l_2 a_1}\right)',$$

where the lowercase letters mean the value observed. Consequently, Table 1 gets mapped to the features shown in Table 2.

Table 2 Feature Values Induced by the Financial Kernel

	a_1/l_1	l_1/a_1	l_2/a_2	a_2/l_2	$a_1l_2/(a_2l_1)$	$l_1a_2/(l_2a_1)$
Firm 1	10/12	12/10	13/11	11/13	[(10)(13)]/[(11)(12)]	[(11)(12)]/[(10)(13)]
Firm 2	4/8	8/4	10/6	6/10	[(4)(10)]/[(6)(8)]	[(6)(8)]/[(4)(10)]

In general, for each year we get all ratios of the form A/L and L/A and year-over-year changes of the form A_1L_2/L_1A_2 and A_2L_1/L_2A_1. Using this kernel, we do not need to create this explicit mapping (i.e., the data in Table 2). This is accomplished by replacing $\langle \mathbf{u}^i, \mathbf{u}^j \rangle$ in Equation (1) with $K_F(\mathbf{u}^i, \mathbf{u}^j)$, or replacing $K(\mathbf{u}^i, \mathbf{u}^j)$ in Equation (2) with $K_F(\mathbf{u}^i, \mathbf{u}^j)$.

In the next section we use the SVM with our financial kernel and apply it to a data set containing fraudulent companies.

6. Data, Testing Methodology, and Results

6.1. Data and Attribute Selection

Fraudulent firms were found using the Securities and Exchange Commission's (SEC's) AAERs (SEC 1995). The first AAER available was #1,190 (issued October 28, 1999) and the last was #2,459 (issued July 11, 2006). There were a total of 1,157 AAERs in the initial sample. Companies merely making mistakes or errors were removed from the data set (fraud requires intention). The SEC rules that apply to fraud are Rule 17(a) from the Securities Exchange Act of 1933, and Rules 13(b)(5), 13b2-1, and 10b-5 from the Securities Exchange Act of 1934. Each of the above statutes relate to fraud as can be ascertained by their descriptions.[2] Eight hundred and ninety-four AAERs included a breach of at least one of the aforementioned SEC rules. Each AAER is for a single company-fraud instance. Conversely, each company-fraud instance can include many AAERs. For each of these, we extracted the company name and the years in which the fraud occurred. We limited our data set to fraud that affected the annual financial statements (10-K). Cases that could not be found in Compustat were dropped. Cases that do not show up on a financial statement via restatement were also dropped (these include such cases as bribery of a foreign official or fraud on a registration statement). We verified that our data set included only companies that had fraudulent financial statements by making sure that there was a restatement for each fraud company-year.

Our final data set included a total of 122 fraudulent firms. Fraud can span a number of years, so we treated each year of fraud as a company-year, ending up with a data set of 205 fraud company-years spanning the years of 1991–2003. We gathered two years of data for each company-year: one for the year of fraud and one for the year prior to fraud.

One-to-one matches (of fraudulent and nonfraudulent firms) are common in the prediction literature (Green and Choi 1997, Summers and Sweeney 1998).

[2] Descriptions can be found at http://www.sec.gov/about/laws/sa33.pdf and http://www.sec.gov/about/laws/sea34.pdf

However, given that fraud occurs less than 1% of the time (Beneish 1999), a one-to-one match is far from mirroring reality. We set out to create a data set that is as close to the observed relative frequencies as possible. Therefore, we only limit our matching criteria by four-digit Standard Industrial Classification (SIC) code and year, allowing there to be many "nonfraud" firms to each fraud firm. We removed any company-year from the nonfraud data set that later restated (including firms that made an error or mistake). The nonfraud data set included a total of 6,427 nonfraudulent company-years. The ratio of fraud to nonfraud firms in our sample is approximately 1:31, or about 3%.[3]

There are no theoretically accepted fraud attributes (Green and Choi 1997). The difficulty in finding fraud attributes is based on the fact that fraud perpetrators are trying to conceal their moves and have a wide variety of financial attributes to work with. Thus, it is difficult to point to a simple set of attributes that encapsulate fraudulent behavior. However, the literature is rich with attempts to use publicly available data to detect fraud. These works generally focus on determining the types of attributes that are correlated with fraud. These attributes can be simple figures from a financial statement or complex combinations of figures, including temporal changes. Utilizing the work of Beneish (1999), Dechow et al. (2009), Summers and Sweeney (1998), and Green and Choi (1997), we gathered a set of 40 attributes that have been used previously in leading publications or in current working papers aimed at fraud detection. These attributes are shown in Table 3 together with their Compustat number (Standard and Poor's 2005) and the paper(s) from which they were referenced.

Every company that the SEC determines to have a financial statement fraud must restate. We gathered the restatement data for the 205 fraud company-years. Table 4 describes our data. Specifically, we report the mean, median, maximum, and minimum values of restatement for the major financial statement accounts. The column labeled "# of firms" shows the number of fraudulent firm-years that restated each account. The restated value for each account represents the firm's adjustment made to correct its accounts. Sales were restated in over half of the fraud cases (106/205). *Earnings per Share* (99/205), *Retained Earnings* (121/205), *Stockholders Equity* (97/205), and *Net Income* (95/205) were also frequently restated. The median restatement value for all accounts was zero. The largest mean among the restated values belonged to *Long-Term Debt* ($1,068,460). Some mean values fall in an unexpected direction, such as *Interest Expense* ($-72,960$), *Working Capital* ($127,090) and *Total Assets* ($108,780). The mean restated value of *Interest Expense* would have been positive, if not for Fannie Mae's huge credit to interest expense. A

[3] During preprocessing, firms are lost because of missing values. The final sample numbers are reported in §6.2.

Table 3 List of Variables (Attributes)

Variable	Reference
Cash and Short-Term Investments (DATA1)	Beneish (1999), Dechow et al. (2009)
Receivables, Total (DATA2)	Beneish (1999), Dechow et al. (2009)
Inventories, Total (DATA3)	Dechow et al. (2009)
*Current Assets, Total** (DATA4)	Beneish (1999), Dechow et al. (2009)
*Current Liabilities, Total** (DATA5)	Beneish (1999), Dechow et al. (2009)
Assets, Total (DATA6)	Summers and Sweeney (1998), Dechow et al. (2009)
*Property, Plant, and Equipment** (DATA7)	Beneish (1999)
Long-Term Debt, Total (DATA9)	Beneish (1999), Dechow et al. (2009)
Sales (DATA12)	Summers and Sweeney (1998), Dechow et al. (2009)
Depreciation and Amortization (DATA14)	Beneish (1999), Dechow et al. (2009)
*Interest Expense** (DATA15)	Summer and Sweeney (1998)
Income Taxes (DATA16)	Summers and Sweeney (1998)
Income Before Extraordinary Items (DATA18)	Dechow et al. (2009)
Price, Calendar Year, Close (DATA24)	Summers and Sweeney (1998)
Common Shares Outstanding (DATA25)	Dechow et al. (2009)
Investments and Advances, Other (DATA32)	Dechow et al. (2009)
Debt in Current Liabilities (DATA34)	Beneish (1999), Dechow et al. (2009)
Retained Earnings (DATA36)	Summers and Sweeney (1998)
Cost of Goods Sold (DATA41)	Beneish (1999), Dechow et al. (2009)
Net Income (DATA172)	(added as a normalizing factor because we use ratios and year-over-year changes in ratios)
Common Equity, Total (DATA60)	Dechow et al. (2009)
*Interest Income** (DATA62)	Summers and Sweeney (1998)
*Receivables, Estimated Doubtful** (DATA67)	Green and Choi (1997)
*Income Taxes Payable** (DATA71)	Beneish (1999)
*Rental Commitments, Min, 1st year** (DATA96)	Dechow et al. (2009)
*Order Backlog** (DATA98)	Dechow et al. (2009)
Depreciation Expense (DATA103)	Beneish (1999)
Preferred Stock, Carrying Value (DATA130)	Dechow et al. (2009)
*Rental Commitments, Min, 2nd year** (DATA164)	Dechow et al. (2009)
*Rental Commitments, Min, 3rd year** (DATA165)	Dechow et al. (2009)
*Rental Commitments, Min, 4th year** (DATA166)	Dechow et al. (2009)
*Rental Commitments, Min, 5th year** (DATA167)	Dechow et al. (2009)
*Deferred Taxes (Income Account)** (DATA50)	Dechow et al. (2009)
Liabilities, Total (DATA181)	Summers and Sweeney (1998)
Selling, General and Admin. Expenses (DATA189)	Beneish (1999)
Short-Term Investments (DATA193)	Dechow et al. (2009)
Price, Fiscal Year, Close (DATA199)	Dechow et al. (2009)
*Financing Activities, Net CF** (DATA313)	Dechow et al. (2009)
*Pension Plans, Anticipated LT ROR on PA** (DATA336)	Dechow et al. (2009)
Employees (DATA29)*	Dechow et al. (2009)

Note. CF, Cash Flow; LT, Long-Term; ROR, Rate of Return; PA, Plan Assets.
*Deleted from final sample because of >25% missing values.

Table 4 Restatement Values by Account

	Maximum	Minimum	Mean	Median	# of firms
	Income statement (in 000s)				
Sales	2,896.91	−38,978.00	−457.00	0.00	106
Cost of Goods Sold	400.00	−9,504.60	−58.08	0.00	83
Selling, General, and Administrative Expenses	248.83	−626.00	−2.05	0.00	66
Depreciation and Amortization	296.00	−787.00	−0.07	0.00	59
Interest Expense	1,181.26	−7,789	−72.96	0.00	42
Income Taxes	25.88	−799.00	−14.38	0.00	75
EPS (*basic, including extraordinary items*)	3.67	−446.48	−4.02	0.00	99
Net Income	176.00	−6,116.50	−66.83	0.00	95
	Balance sheet (in 000s)				
Minority Interest	2.24	−3,219.00	−53.69	0.00	22
Working Capital	18,555.01	−613.80	127.09	0.00	77
Capital Expenses	5,305.80	−61.61	41.06	0.00	51
Property, Plant, and Equipment	4,970.90	−949.86	25.77	0.00	41
Total Assets	17,508.98	−1,369.38	108.78	0.00	94
Long-Term Debt	139,079	−532.80	1,068.46	0.00	41
Retained Earnings	4,459.00	−10,289.38	−55.32	0.00	121
Stockholders Equity	15,611.00	3,088.68	27.69	0.00	97

Note. EPS, Earnings per Share.

large portion of the positive *Working Capital* mean was based on Adelphia's restatement.[4] Without Adelphia and Fannie Mae, the mean *Total Asset* restatement would have been negative. The anomalies encountered in this analysis of fraud firms underscore the complexity of the fraud detection task.

Preprocessing included three steps: (i) transformation of attributes with a zero value, (ii) removing attributes with a large number of missing values, and (iii) removing firms with missing values. Quantitative attributes with a value of zero are a problem because the FK's mapping will result in division by zero. To avoid this problem, zero values are replaced with 0.0001. We remove firms with missing values as is done in the overwhelming majority of empirical accounting research papers. Two recent examples are Khan and Watts (2009) and Callen et al. (2009). In the case of fraud detection, all of the papers that we compare to (in §6.3 below) that use publicly available data delete firms with missing values. Attributes with greater than 25% missing values were removed, yielding 23 final attributes. The 17 deleted attributes are marked with asterisks in Table 3.

[4] The fact that *Working Capital* is restated positively on fraud cases is intuitive as much of fraud is reporting nonexistent sales. The reversal of these sales often entails putting inventory back onto the balance sheet, thus raising *Working Capital*. The sign of *Cost of Goods Sold* indicates a drop upon restatement. This is no surprise either because *Cost of Goods Sold* is tied directly to sales, so a drop in sales is a drop in *Cost of Goods Sold*.

Table 5 SVM-FK Results on the Test Set

C_{+1}: C_{-1} weightings (fraud:nonfraud)	Fraud recall (Type I error) (%)	Nonfraud recall (Type II error) (%)	AUC
1:1	0.00 (100)	100 (0)	0.816
5:1	20.0 (80)	93.4 (6.6)	0.726
10:1	32.0 (68)	87.2 (12.8)	0.739
15:1	44.0 (56)	86.2 (13.8)	0.774
20:1	56.0 (44)	86.3 (13.7)	0.810
50:1	80.0 (20)	90.6 (9.4)	0.878
100:1	72.0 (28)	89.1 (10.9)	0.877
150:1	80.0 (20)	84.7 (15.3)	0.878
200:1	80.0 (20)	84.4 (15.6)	0.880

In the next section we report the results of our experiments.

6.2. Testing and Results

We validate the SVM-FK methodology by training on the early years of the data set and testing on future years. By training on early years and testing on future years, we are matching the situation faced by an investor or auditor who needs to make judgments on new, unseen data with only prior data for support. Our training sample includes data from 1991 to 2000, and our test data set includes data from 2001 to 2003. The training sample, after preprocessing, includes 107 fraud company-years and 2,205 nonfraud company-years. The holdout sample includes 25 fraud company-years and 982 nonfraud company-years.

After deleting attributes with greater than 25% missing values, we have 23 attributes from prior research to use, resulting in 1,518 FK features. We add a numerical value for the year as a control. In Table 5 we report the results of our test sample. The results show the recall for both classes (along with Type I and Type II errors) for different C_{+1}:C_{-1} weightings (fraud:nonfraud) in the SVM objective. It is likely that a user of this type of model would weight the risk of a Type I error much higher than the risk of a Type II error. Beneish (1999) conjectures that the right cost ratio for investors is between 20:1 and 30:1. We also report the area under curve (AUC) corresponding to our results. The AUC is the area under the receiver operating characteristic (ROC) curve[5] (Fawcett 2006).

Table 5 shows that the SVM-FK is able to achieve 80% recall on fraud while achieving 90.6% recall on nonfraud at the 50:1 cost ratio. As the cost ratio increases above 50:1, the fraud results stay the same but nonfraud recall gradually deteriorates. The AUC is 0.878 (an ROC curve comparing our results with

[5] AUC offers information as to the appropriateness of a model. AUC is based on the rate at which true positives are found compared to the rate of false positives during the prediction process (i.e., creating an ROC curve requires ranking the predictions and picking the highest ranked prediction first).

Table 6 Comparison of Results with Previous Fraud Detection Methods

Author(s)	Method	Recall: Percentage correct—(sample size)	Results on training or test set	Type of data
Loebbecke et al. (1989)	Assessment model	86—Fraud (77)	Training	Nonpublic data
Hansen et al. (1996)	Qualitative response model	56—Fraud (77) 90—Nonfraud (305)	Test	Nonpublic data
Bell and Carcello (2000)	Logistic regression	81—Fraud (77) 86—Nonfraud (305)	Test	Nonpublic data
Beneish (1999)	Probit	56—Fraud (74) 90.2—Nonfraud (2,332)	Test	Publicly available data
Summers and Sweeney (1998)	Logistic regression	67—Fraud (51) ?— Nonfraud (51)	Training	Publicly available data
Green and Choi (1997)	Neural network	74—Fraud (46) 68.4—Nonfraud (49)	Test	Publicly available data
Dechow et al. (2009)	Logistic regression	64.5—Fraud (293) 66.35—Nonfraud (79,358)	Test	Publicly available data
This paper	SVM-FK	80.0—Fraud (132) 90.6—Nonfraud (3,187)	Test	Publicly available data

the results of other studies on these data is shown in the next section). The novelty of our research is the nonlinear mapping of the attributes by the financial kernel into relevant features combined with the structural risk minimization offered by the SVM. In the following section we compare our results to other leading fraud detection research.

6.3. Comparison with Prior Research Methods

For comparison purposes, the prediction results of other leading fraud research is shown below in Table 6, together with the size of the data sets (the larger, more comprehensive data sets indicate a higher likelihood of generalization to the population) and the type of data used (nonpublic or publicly available). Nonpublic data is much more costly and time consuming to acquire. Also, it is impossible to collect unless one has a relationship with the firm.

When compared against other research using publicly available data, we report the highest accuracy on fraud cases, and we have the second-largest data set. The highest overall results on fraud cases are found in the seminal work by Loebbecke et al. (1989). However, the authors do not test their model on nonfraud cases (therefore, there is no possibility of Type II errors). The comparison in Table 6 is meant to frame our research with respect to the current fraud detection research stream. Because the data sets are not shared, the comparisons between the methods are necessarily qualitative. Some important differences are readily noted including testing method, sample size, and skew of fraud and nonfraud data. We briefly discuss the testing methods of the other papers to aid in the comparison. In Loebbecke et al. (1989), the assessment model is created, modified, and tested on the same data. There are no nonfraud examples, so the assessment model is not tested for Type II errors. Bell and Carcello (2000), using private data, estimate a logistic regression model hundreds of times on random training samples, looking ahead at the test results to determine the best model. Summers and Sweeney (1998) use a cascaded logit model with fraud as the dependent variable. The results they report are based on a single sample, with no holdout set. Therefore, it is not possible to determine whether their model has predictive power. Green and Choi (1997) have a limited sample size. Hansen et al. (1996), Beneish (1999), Bell and Carcello (2000), Dechow et al. (2009), and our paper use similar testing methodologies. All five have training and holdout sets and report the results for both fraud and nonfraud. Only our paper, Dechow et al. (2009), and Beneish (1999) test the models on future data (using estimation data from prior years).

To directly compare the efficacies of the research methods (which include the methodologies and the attributes), we replicate the studies that use publicly available data using our data set. These studies include Green and Choi (1997), Beneish (1999), and Dechow et al. (2009).[6] We make every attempt to faithfully replicate their testing methodologies by gathering the same financial data they used for their studies and tuning their models exactly as their papers suggested. The details of each replication are explained below.

6.3.1. Green and Choi (1997). All variables were Winsorized at the 1st and 99th percentiles. All variables were transformed via a simple percentage

[6] Summers and Sweeney (1998) used variables (insider trading statistics) that were key to their study. These insider trading variables were received from the SEC at the time of their study but are no longer available.

change (SPC)[7] and scaled by a linear scaling method. Details of the SPC and the scaling method can be found in Green and Choi (1997). The NN was set up with the following parameters to match the authors work: The network is a back propagation network with eight input nodes, four hidden nodes, and either one or two output nodes.[8] The learning rate and momentum were both set to 0.1, and the learning epochs were limited to 10,000. The Green and Choi (1997) variables are as follows: *Allowance for Doubtful Accounts/Net Sales*, *Allowance for Doubtful Accounts/Accounts Receivable*, *Net Sales/Accounts Receivable*, *Gross Margin/Net Sales*, *Accounts Receivable/Total Assets*, *Net Sales*, *Accounts Receivable*, *Allowance for Doubtful Accounts.*

6.3.2. Dechow et al. (2009). All variables were Winsorized at the 1st and 99th percentiles. We used SAS Proc Logistic to develop the logistic regression model. We also used SAS Proc Logistic to test the model on the test set. Starting with the predicted value from the logistic regression function, Dechow et al. (2009) utilized the following transformations to achieve binary classification:

$$Probability = \frac{e^{(PredictedValue)}}{(1+e^{(PredictedValue)})}$$

$$UnconditionalProbability = TotalFraudFirms/(TotalFraudFirms + TotalNonfraudFirms)$$

$$F\text{-Score} = Probability/UnconditionalProbability.$$

An F-score of 1.00 means that the firm has the same probability of fraud as unconditional expectation (Dechow et al. 2009). The threshold becomes 1.00 (following Dechow et al. 2009) such that $F > 1$ labels the firm fraudulent and $F \leq 1$ labels the firms nonfraudulent. We summarize the Dechow et al. (2009) variables as follows: *RSST Accruals*, *Change in Receivables*, *Change in Inventory*, *Change in Cash Sales*, *Change in Earnings*, *Actual Issuance*, *Abnormal Change in Employees*, *Existence of Operating Leases*, *Book to Market*, *Lagged Market-Adjusted Stock Return*.

A detailed explanation of these variables can be found in Dechow et al. (2009).

[7] This was one of three transformations in Green and Choi (1997). We chose this one because it was used to achieve the highest accuracy on their test set.

[8] Green and Choi (1997) use one output node and make the threshold between fraud and nonfraud 0.5 (nonfraud is labeled 1 and fraud is labeled 0). We develop our NN in Weka using this method as well as setting up two output nodes, one for fraud and one for nonfraud, using a nominal class variable. Our results are not affected by these design choices. We optimize the NN threshold for performance after the fact in an attempt to improve results.

Table 7 Comparative Results Using the Same Data Set

Author(s)	Method	Final sample size (after excluding firms with missing values)	Percentage correct	AUC
Beneish (1999)	Probit	149 Fraud 3,389 Nonfraud	54.2 Fraud 45.5 Nonfraud	0.492
Green and Choi (1997)	Neural network	192 Fraud 3,173 Nonfraud	100.0 Fraud 7.1 Nonfraud	0.472
Dechow et al. (2009)	Logistic regression	57 Fraud 1,244 Nonfraud	70.0 Fraud 84.9 Nonfraud	0.762
This paper	SVM-FK	132 Fraud 3,187 Nonfraud	80.0 Fraud 90.6 Nonfraud	0.878

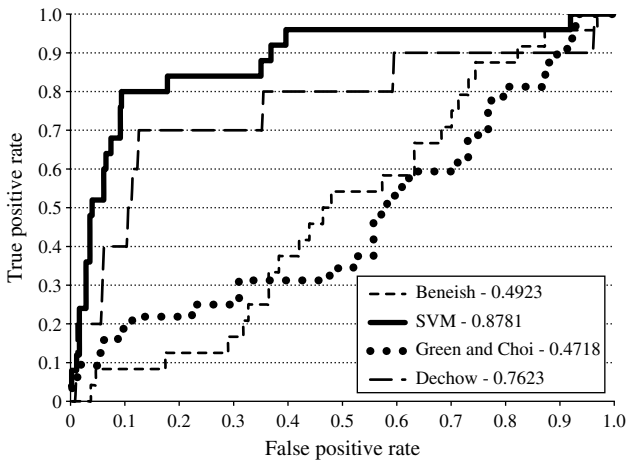
6.3.3. Beneish (1999). All variables were Winsorized at the 1st and 99th percentiles. We used SAS Proc Logistic (link = probit) to develop the probit model on the training set. We also used SAS Proc Logistic to test the trained model on the holdout sample. We used the Beneish (1999) method of picking the best threshold by choosing the one that gives the lowest expected cost of misclassification (ECM), where ECM is defined as ECM $= P(M)P_{\mathrm{I}}C_{\mathrm{I}} + [1 - P(M)]P_{\mathrm{II}}C_{\mathrm{II}}$, where $P(M)$ is the prior probability of encountering manipulators, P_{I} and P_{II} are the conditional probabilities of Type I and Type II errors, respectively, and C_{I} and C_{II} are the cost of Type I and Type II errors, respectively. The ratio of the number of fraud cases to the number of cases in the entire sample is denoted by $P(M)$. We summarize the Beneish (1999) variables as follows: *Change in Receivables* (deflated by sales), *Change in Gross Margin* (deflated by sales), *Ratio of Noncurrent Assets* (less *Property, Plant, and Equipment*) *to Total Assets*, *Change in Sales*, *Change in Depreciation*, *Change in Selling, General and Administrative Expenses* (deflated by sales), *Change in Liabilities* (deflated by *Total Assets*), and *Total Accruals* as a percentage of *Total Assets.*

A detailed explanation of these variables can be found in Beneish (1999).

The results of our replications are shown in Table 7 below. Beneish (1999) was able to get both fraud and nonfraud correct about 50% of the time while optimizing the training threshold (as is done in his study) at the 30:1 cost ratio.[9] The Green and Choi (1997) method labels everything as nonfraud. To check if this is a threshold problem, we use Beneish's (1999) function (ECM) to optimize the threshold (looking ahead). The tabulated results report the best cost ratio as 100% accuracy on fraud and 7.1% accuracy on nonfraud

[9] When moving to the 40:1 cost ratio, the fraud recall improves to 87.5% but the nonfraud recall drops to 20.2%. At the 20:1 cost ratio, the fraud recall drops to 45.8%, whereas the nonfraud recall becomes 56.1%.

Figure 1 ROC Curve Comparing Research Methodologies



using a 30:1 cost ratio.[10] The Dechow et al. (2009) model was able to classify 70.0% of fraud firms correctly and was also able to classify 84.9% of nonfraud firms correctly. The SVM-FK was able to classify 80% of fraud firms correctly and 90.6% of nonfraud firms correctly. The SVM-FK was superior in both types of recall.

We also report the AUC in Table 7 and the comparative ROC curve in Figure 1. The Beneish (1999) and Green and Choi (1997) models have similar AUCs (0.492 and 0.472, respectively). The Dechow et al. (2009) model has a much higher AUC at 0.762. The highest AUC is achieved by the SVM-FK at 0.878. Based on the curves seen in Figure 1, it is clear that the Dechow et al. (2009) model and SVM-FK are better at predicting and recalling fraud cases than the Beneish (1999) and Green and Choi (1997) models. Given a set of companies, the SVM-FK picks 80% of the fraud cases while making few mistakes. The SVM-FK ROC curve dominates that of Dechow et al. (2009). This suggests that if a sample of companies is to be picked for auditing using these models, the SVM-FK is more likely to select more of the true fraud cases compared to competing methods. This behavior is quite valuable for budget- and time-constrained auditors or regulators, whose job it is to pick likely fraud cases for further investigation with minimal false positives. Auditors may be liable for missing fraud cases, making true positives extremely important. However, economic constraints limit further investigation, because false positives are costly. In the case of a regulatory body, false positives waste scarce investigative labor, raising the probability that a true positive will go uninvestigated.

[10] When moving to a 20:1 cost ratio, the fraud recall becomes 21.9% and nonfraud recall becomes 88.2%.

Why does the SVM-FK detect fraud better? The answer likely lies in the fact that the SVM-FK combines domain knowledge (via the kernel) with a machine learning method that addresses generalization directly. The NN solution of Green and Choi (1997) was focused on introducing NNs to the fraud detection problem, thus highlighting the novelty of the methodology. Beneish (1999) and Dechow et al. (2009) showcased human expertise by developing ratios based on the continually evolving understanding of financial statement accounts that are commonly manipulated. Both of these papers move forward the human expertise in variable selection.

Our method utilizes the expert knowledge from these studies by incorporating their financial attributes and ratios. Unlike those methods where ratios are developed and tested manually, the kernel allows us to create and inject ratios and others (i.e., features) into the learning algorithm without having to develop and test them manually. The SVM, then, determines the weights associated with these ratios. The SVM includes nonlinear capabilities (via the kernel), whereas NNs attain these through network structure and activation functions. However, unlike NNs the resulting SVM function is transparent, making it amenable to analysis, as in logit or probit. Also, the SVM is designed to ameliorate overfitting by trading off generalization ability with training error. Logit, probit, and NN models do not perform such tradeoffs. Therefore, they are more likely to overfit as the number of features increase.

Furthermore, our method makes very few assumptions compared to the other methods. We start with the raw data, pull out variables with many missing values, leave out firms with missing values (as virtually all the studies do), replace 0 values with 0.0001 values, and then run the data through the FK. Our only look-ahead bias comes from training with various cost-sensitive weights. Beneish (1999) Winsorizes the raw data, develops a probit function with a separate thresholding cutoff for each weight in the estimation sample, and determines the best results based on a postprocessing function (ECM) that uses the weights and the errors on the training sample. ECM incorporates the skew of the data sample to find a cutoff based on the sample parameters. Dechow et al. (2009) Winsorize the raw data and use the F-score as a postprocessing transformation after logit. The F-score is a method (like ECM) of incorporating the probability of fraud based on the skew in the sample size. Green and Choi (1997) use two data transformations before processing through the NN.

6.4. Additional Analyses

To study what features might be important in this data set, we ranked the features from largest to smallest based on $|w_j \bar{f}_j|$, where w_j is the jth component of

Table 8 Top Five Features of SVM-FK

Feature	Weight (absolute value)	Ratio	Year	Correlation with fraud
1	0.403	*Sales/Preferred Stock, Carrying Value*	$t-1$	Positive
2	0.384	*Selling, General, and Administrative Expenses/Investments and Advances, Other*	t	Negative
3	0.275	*Total Assets/Investments and Advances, Other*	$t-1$	Positive
4	0.273	*Sales/Investments and Advances, Other*	$t-1$	Negative
5	0.245	*Total Assets/Short-Term Investments*	t	Positive

the SVM $\mathbf{w}$ vector and $\bar{f}_j$ is the average of the jth feature resulting from the FK. The top five features are shown in Table 8.

In each of the top five features, the denominator is a part of RSST accruals. RSST accruals are utilized in the Dechow et al. (2009) study as indicators of fraudulent behavior. Using these components as deflators has not been attempted in previous research. The traditional deflators are total assets and sales. Durtschi and Easton (2005) find that the detection of earnings management can be affected by choice of deflator. We are not limited by this assumption because we make no a priori assumption about how each variable can best be used. As *Sales* in year 1 deflated by *Preferred Stock* in year 1 increases, so does the chance of fraud. Ratio 4 shows that as *Sales* in year 1 (deflated by *Other Investments and Advances*) increase, the chance of fraud decreases. This suggests the commonly assumed fact that fraud firms are under increased economic pressure because poorer performance (in this case, lower sales). Furthermore, a fraud firm is likely to have lower *Preferred Stock* and higher *Other Investment and Advances* than a nonfraud firm in the year prior to fraud. Ratio 3 also has *Other Investments and Advances* in year 1 as a denominator. With *Total Assets* as a numerator, one would expect this to be negatively correlated with fraud. However, this is not the case. *Total Assets* in year 2 is also a numerator in Ratio 5 and is skewed to fraud. This goes against the common intuition that fraud companies are smaller (as size is often judged by *Total Assets*). Our sample includes some very important large fraud companies, including Enron and WorldCom. However, it is not *Total Assets* by itself that is associated with fraud but the relationship between *Total Assets* and *Other Investments and Advances* and *Short-Term Investments*. These relationships would be difficult to predict, even by fraud experts.

Notice that the numerators in all cases are very basic financial statement items. The denominators are more detailed financial information. The top five features are but a small part of the total function (which includes 1,518 features), so placing too much importance on these five is likely ill advised. These features alone do not explain the power of the method. The power comes from the combination of all the features. With the SVM-FK, the researcher benefits from getting many features (some of which are likely unknown a priori). Furthermore, unlike NNs, the researcher can analyze the important features after the fact.

Those who commit fraud try to hide it by gaming the system. Most professionals engaged in fraud know what auditors look for when they suspect fraud. To this end we point to a few interesting facts. The top features we find are quite different than those of earlier researchers such as Beneish et al. (1999) and Green and Choi (1997), whose data spanned from 1982 to 1992 and from 1982 to 1990, respectively. In addition, we showed that their approach does not lead to satisfactory results in this data set, whereas the approach of Dechow et al. (2009), which also uses RSST related variables, worked very well. We think that the discrepancies in results are due more to features used rather than specific induction technique that was used. We also believe that this difference, at least partially, can be explained by changes in fraud tactics over the years. Therefore, a method that utilizes exhaustive combinations of potential fraud variables has a better chance of catching fraud effectively than methods that restrict themselves to a few possible constructs. SVM is theoretically able to handle an infinite number of features. FK is a mapping that allows for an exhaustive combination of ratios and year-over-year changes. For a complex, moving target like fraud, this combination may be what's necessary for effective prediction.

To use this as a support tool in a professional setting one must not only consider the recall (as was done above) but also the precision. Precision is the number of frauds divided by the total number of firms counted as fraud. In a real-world case with a skewed data set, the fraud precision is necessarily small (because there are so many fewer fraud cases, even if one gets a high percentage of nonfrauds correct, the nonfraud errors will greatly outnumber the correct fraud cases). An effective predictor will pick the frauds without labeling many nonfrauds as frauds. From the perspective of a professional, each nonfraud labeled as fraud carries a cost. As an auditor or regulator it takes time and effort to investigate the firms that are flagged as potentially fraudulent. Investment professionals risk losing out on potentially high return investments if they are erroneously flagged as fraudulent. The Dechow et al. (2009) model results in a 2.25% fraud precision. This means that for every fraud company that is correctly labeled as a

fraud company, approximately 44 nonfraud firms are erroneously labeled as fraudulent. The SVM-FK has a 17.86% fraud precision. This means that for every fraud company that is correctly labeled as a fraud company, approximately six nonfraud firms are erroneously labeled as fraudulent.

7. Conclusion and Future Research

This paper developed a methodology for detecting management fraud. A domain-specific kernel, the financial kernel, was created that implicitly maps relevant financial attributes to ratios and year-over-year changes of the ratios. Careful consideration was given in constructing the features necessary to detect management fraud. The kernel is used as a component in the SVM.

A data set of fraudulent companies was gathered using accounting and auditing enforcement releases. The fraud firms were matched with nonfraud firms based on a four-digit SIC code and year. To reflect reality, the data set was skewed allowing for many nonfraud firm-years for each fraud firm-year. The resulting data set included 137 fraudulent firm years and 3,187 nonfraudulent firm years. The attributes used for mapping into FK were taken from the leading previous fraud detection research, putting emphasis on the model for discriminating between fraud and nonfraud cases in a higher dimensional feature space. The estimation set was created using fraud and nonfraud firm years from 1991 to 2000, and the holdout set was created using fraud and nonfraud firms from 2001 to 2003. Training on early years with a later year holdout set enabled us to test the prediction ability of the model. The results show that the SVM-FK correctly labels 80.0% of the fraud cases on a holdout set while also correctly labeling 90.6% of the nonfraud cases with the same model. The result is an improvement on prior research using only publicly available data and is competitive with the best research using nonpublic data (as shown in Tables 6 and 7). This methodology has the potential to assist regulators, auditors and investors as a tool for determining which firms are at higher risk for fraud. No machine learning tool can be relied upon solely, but the SVM-FK can be utilized in conjunction with expert analysis and judgment to shorten the search for the right candidates for investigation for regulatory agencies. An audit firm could use the methodology to support its decisions on "gray area" firms, which may need further testing for fraud and other irregularities. An investor can use the methodology to find the firms to avoid and lower the risk of encountering the major drop in share price that accompanies the findings of fraud.

Much of fraud research has focused on finding the variables that are highly associated with fraud and analyzing those variables for insight. This paper develops a methodology that can be used with the variables from extant literature to develop models that create novel combinations of these variables to aid in the prediction of fraud. We believe that future studies should take into account the fact that fraudsters will change their tactics to hide fraud. An inductive principle that anticipates this strategic behavior might be more suitable to get ahead in this arms race, one in which the fraudsters are currently a step ahead.

A natural extension of this work would be the development of a trading rule based on the results of the SVM-FK. Beneish (1997) supplies the framework for such an extension. This research can also be extended by improving the methodology to incorporate nonfinancial information that is germane to the domain of fraud detection. More generally, the methodology can be combined with extant research from other domains to predict different phenomena in accounting, such as bankruptcy, corporate restatements, abnormal returns, and firms likely to be subject (object) of mergers and acquisitions to name a few.

Acknowledgments

The authors thank Kristin Bennett, Joe Ecker, Karl Hackenbrack, Scott Jackson, Al Leitch, Tom Lopez, Gary McGill, participants at Rensselaer Polytechnic Institute research seminar, and two anonymous reviewers for providing insightful comments and suggestions. They also thank Leigh Salzsieder for providing research assistance.

Appendix

FK Construction

The financial kernel is constructed using a graph kernel (Takimoto and Warmuth 2003) on the financial attributes followed by a normalization kernel (Graf and Borer 2001) applied to the resulting features plus the *Year* value. The normalization kernel merely normalizes the features resulting from the graph kernel (for further information on the normalization kernel, see Graf and Borer 2001). This sequence results in a proper kernel. The application of the graph kernel is further explained below.

Consider a graph $G(A, E)$ with nodes a and edges e (see Figure A.1). All edges e in this graph are base kernels (for example, a polynomial kernel on a component of the attribute space). To differentiate base kernels from general kernels, the base kernels are denoted as $K(u_i, v)$, where the

Figure A.1 Graph Kernel

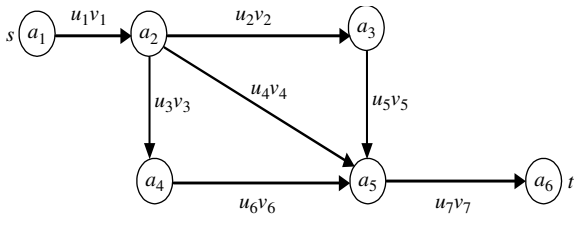
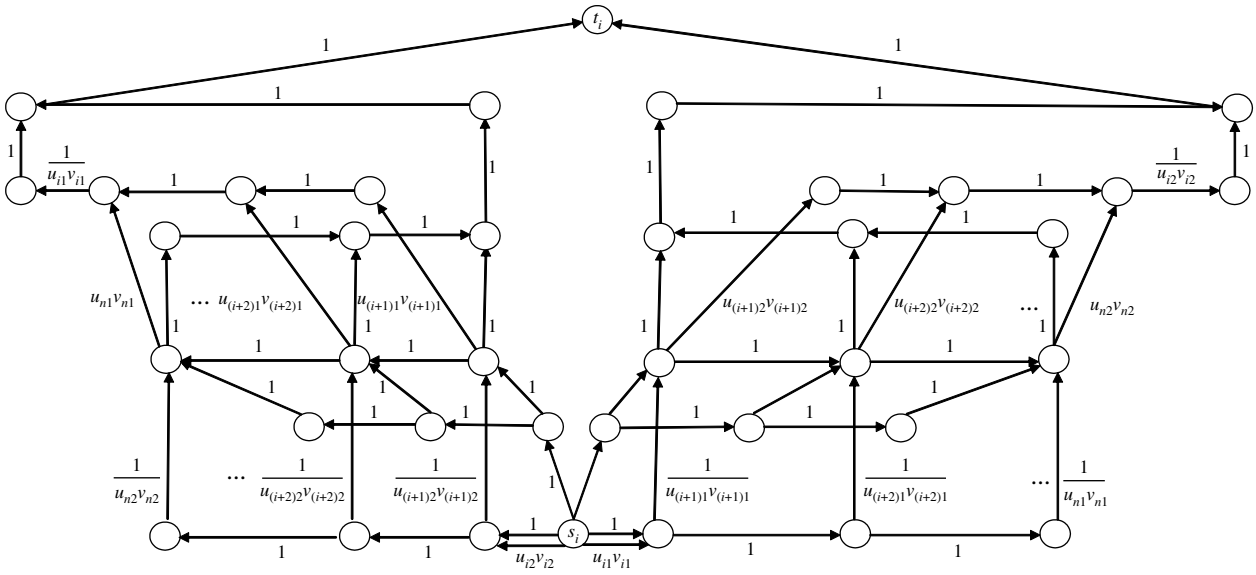


Figure A.2 Financial Kernel



unbolded u_i and v_i are scalars. For this example, each edge e has a kernel of the following form: $K(u_i, v_i) = \langle u_i, v_i \rangle$. Any path from s to t yields a feature. The feature is arrived at via the product of all edges in the path between s and t. If $s = a_1$ and $t = a_2$, then there would be a single feature, u_1. If $s = a_1$ and $t = a_3$, there would also be a single feature, u_1u_2, but the feature would be the product of the two base kernels on the path $p = (a_1a_2a_3)$. Three paths converge at node a_5, specifically, $p_1 = (a_1a_2a_3a_5)$, $p_2 = (a_1a_2a_5)$, and $p_3 = (a_1a_2a_4a_5)$. Node a_5 can be seen as a kernel that sums the products of the base kernels on each path. If Node a_5 were t, the output would be the sum of all paths into node a_5, $p_1 + p_2 + p_3$ or $u_1u_2u_5 + u_1u_4 + u_1u_3u_6$. In general, at each node a (except s), all paths from s to a are summed. The contribution of a path to the kernel is based on the product of its edges. In general, all paths from s to t create features. This allows the researcher to create her own kernel by choosing the structure of the graph.

Formally, the graph kernel is a directed graph G with a source vertex s of in-degree 0 and a sink vertex t of out degree 0. A directed graph is one where the flow on each edge is in a single direction. Each edge is labeled with a base kernel. It is assumed that this is a simple graph, meaning that there are no directed loops. In general, loops are allowed, but that makes it difficult to prove that the resulting mapping is, indeed, a kernel. Takimoto and Warmuth (2003) proved that a directed, acyclical graph with base kernels on the edges is indeed a kernel.

Shawe-Taylor and Cristianini (2004) describe the graph kernel as follows: Let P_{st} be the set of directed paths from s to t where a path $p = (a_0a_1 \cdots a_d)$. The product of the kernels associated with the edges of p is $K_p(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{d} K_{(a_{i-1} \to a)}(\mathbf{u}, \mathbf{v})$, where $a_{i-1} \to a_i$ represents the i-1st edge on path p.

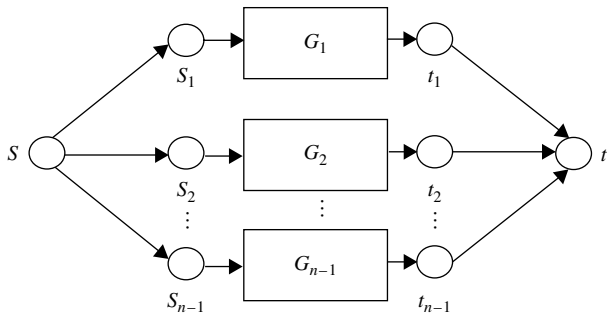
The graph kernel is the aggregation of all $K_P(\mathbf{u}, \mathbf{v})$ and can be seen as follows:

$$K_G(\mathbf{u}, \mathbf{v}) = \sum_{p \in P_{st}} K_P(\mathbf{u}, \mathbf{v}) = \sum_{p \in P_{st}} \prod_{i=1}^{d} K_{(a_{i-1} \to a_i)}(\mathbf{u}, \mathbf{v}).$$

Using these ideas, we now construct the financial kernel, $K_F(\mathbf{u}, \mathbf{v})$, which is a directed graph $G \in (A, E)$ with base kernels on all edges e and $K(\mathbf{u}, \mathbf{v})$ on a. The financial kernel has as input n attributes per year for two years. The attributes vector is $\mathbf{u} = (u_{11}, \ldots, u_{n1}, u_{12}, \ldots, u_{n2})'$. See Figures A.2 and A.3 for illustrations of the financial domain kernel. Figure A.2 illustrates one of the $n-1$ graphs that make up the financial kernel. Each of the $n-1$ graphs has a source node s_i and a sink node t_i. The graphs decrease in size with n. The reason is that each graph carries information for attributes i through n. Each path from source to sink is a feature. The number of features is equal to the number of paths. All $n-1$ graphs from Figure A.2 are brought together by the graph in Figure A.3. The paths from s to t make up all of the features in $K_F(\mathbf{u}, \mathbf{v})$.

We can have as many different kernels as there are edges. For the creation of a financial kernel, we limited the base kernels to two forms. The first one is the standard inner product kernel of $K(u_i, v_i) = \langle u_i, v_i \rangle$, and the second one is $\tilde{K}(u_i, v_i) = 1/u_iv_i$. According to Takimoto and Warmuth (2003), to prove that $K_F(\mathbf{u}, \mathbf{v})$ is a kernel, we need only have a directed graph without cycles and show that each edge e is a valid kernel. Examination of Figures A.2 and A.3 clearly shows that the graph is directed and free of cycles. We

Figure A.3 Financial Kernel Aggregation



need to show that both $K(u_i, v_i)$ and $\tilde{K}(u_i, v_i)$ are kernels; $K(u_i, v_i)$ is simply the standard inner product kernel, and $\tilde{K}(u_i, v_i)$ can be shown to be a kernel as follows. Let $f(u_i) = u_i^{-1}$, $i = 1 \cdots n$, and let $f(v_i) = v_i^{-1}$, $i = 1 \cdots n$. By Cristianini and Shawe-Taylor (2000, p. 42), $\tilde{K}(u_i, v_i) = f(u_i)f(v_i)$ is a kernel.

References

Abbot, L. J., S. Parker, G. Peters. 2004. Audit committee characteristics and restatements. *Auditing* **23**(March) 69–77.

Agresti, A. 1990. *Categorical Data Analysis*. John Wiley & Sons, New York.

American Institute of Certified Public Accountants (AICPA). 2002. What does new audit standard SAS no. 99, *consideration of fraud in a financial statement audit*, mean for business and industry members? *The CPA Letter* (November), http://www.aicpa.org/pubs/cpaltr/nov2002/supps/busind1.htm.

Asare, S. K., A. M. Wright. 2004. The effectiveness of alternative risk assessment and program planning tools in a fraud setting. *Contemporary Accounting Res.* **21**(2) 325–352.

Bell, T. B., J. V. Carcello. 2000. Research notes, a decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: J. Practice Theory* **19**(1) 169–175.

Beneish, M. 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *J. Accounting Public Policy* **16**(3) 271–309.

Beneish, M. 1999. The detection of earnings manipulation. *Financial Analysts J.* **55**(5) 24–36.

Callen, J. L., J. Livnat, D. Segal. 2009. The impact of earnings on the pricing of credit default swaps. *Accounting Rev.* **84**(5) 1363–1394.

Cristianini, N., J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Cristianini, N., J. Shawe-Taylor, H. Lodhi. 2002. Latent semantic kernels. *J. Intelligent Inform. Systems* **18**(2–3) 127–152.

Dechow, P. M., W. Ge, C. R. Larson, R. G. Sloan. 2009. Predicting material accounting misstatements. AAA 2008 Financial Accounting and Reporting Section (FARS) Paper. http://ssrn.com/abstract=997483.

Durtschi, C., P. Easton. 2005. Earnings management? The shapes of the frequency distributions of earnings metrics are not evidence ipso facto. *J. Accounting Res.* **43**(4) 557–592.

Eisenbeis, R. 1987. Discussion, supplement to Srinivasan, V. and Kim, Y. H. Credit granting: A comparative analysis of classification procedures. *J. Finance* **42**(3) 681–683.

Fanning, K., K. O. Cogger, R. Srivastava. 1995. Detection of management fraud: A neural network approach. *Proc. 11th Conf. Artificial Intelligence Appl.*, IEEE Computer Society, Washington, DC, 220–223.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Lett.* **27**(8) 861–874.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**(7) 179–188.

Francis, J., R. LaFond, P. Olsson, K. Schipper. 2005. The market pricing of accruals quality. *J. Accounting Econom.* **39**(2) 295–327.

Genton, M. G. 2001. Classes of kernels for machine-learning: A statistics perspective. *J. Machine Learn. Res.* **2**(12) 299–312.

Graf, A. B. A., S. Borer. 2001. Normalization in support vector machines. *Proc. 23rd DAGM-Sympos. Pattern Recognition*, Springer-Verlag, London, 277–282.

Green, P., J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing: J. Practice Theory* **16**(1) 14–29.

Hackenbrack, K. 1993. The effect of experience with different sized clients on auditor evaluations of fraudulent financial reporting indicators. *Auditing: J. Practice Theory* **12**(1) 99–100.

Hansen, J. V., J. B. McDonald, W. F. Messier, T. B. Bell. 1996. A generalized qualitative-response model and the analysis of management fraud. *Management Sci.* **42**(7) 1022–1033.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Proc. 10th European Conf. Machine Learning*, Springer-Verlag, London, 137–142.

Khan, M., R. L. Watts. 2009. Estimation and empirical properties of a firm-year measure of accounting conservatism. *J. Accounting Econom.* **48**(2–3) 132–150.

Loebbecke, J. K., M. M. Eining, J. J. Willingham. 1989. Auditors' experience with material irregularities: Frequency, nature, and detectability. *Auditing: J. Practice Theory* **9**(1) 1–28.

McNichols, M., P. Wilson. 1988. Evidence of earnings management from the provision for bad debts. *J. Accounting Res.* **26** 1–31.

Messier, W. F., J. V. Hansen. 1988. Inducing rules for expert system development: An example using default bankruptcy data. *Management Sci.* **34**(12) 1403–1416.

New York Stock Exchange. 2003. Final NYSE corporate governance rules. Report, http://www.nyse.com/pdfs/finalcorpgovrules.pdf.

Pincus, K. V. 1989. The efficacy of a red flags questionnaire for assessing the possibility of fraud. *Accounting, Organ. Soc.* **14**(1–2) 153–163.

Quinlan, J. R. 1996. Decision trees and instance-based classifiers. A. B. Tucker, ed. *CRC Handbook of Computer Science and Engineering*. CRC Press, Boca Raton, FL, 521–535.

Ragothaman, S., J. Carpenter, T. Buttars. 1995. Using rule induction for knowledge acquisition: An expert systems approach to evaluating material errors and irregularities. *Expert Systems with Appl.* **9**(4) 483–490.

Rüping, S. 2001. SVM kernels for time series analysis. R. Klinkenberg, S. Rüping, A. Fick, N. Henze, C. Herzog, R. Molitor, O. Schrüder, eds. *LLWA 01—Tagungsband der GI-Workshop-Woche Lernen—Lehren—Wissen—Adaptivität, Dortmund, Germany*, 43–50.

Securities and Exchange Commission (SEC). 1995. Selected accounting and auditing enforcement releases. http://www.sec.gov/divisions/enforce/friactions.shtml.

Shawe-Taylor, J., N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

Standard & Poor's. Compustat database. 2005. Accessed June 2009, http://www.compustat.com/.

Summers, S. L., J. T. Sweeney. 1998. Fraudulentyl misstated financial statements and insider trading: An empirical analysis. *Accounting Rev.* **73**(1) 131–146.

Takimoto, E., M. Warmuth. 2003. Path kernels and multiplicative updates. *J. Machine Learning Res.* **4** 773–818.

Tam, K., M. Kiang. 1992. Managerial applications of neural networks: The case of bank failure predictions. *Management Sci.* **38**(7) 926–947.

Tsai, L., G. Koehler. 1993. The accuracy of concepts learned from induction. *Decision Support Systems* **10**(2) 161–172.

U.S. Congress. *Sarbanes-Oxley Act of 2002*. HR 3763. 107th Cong., 2nd Sess. Pub. L. 107-204, 116 Stat. 745 (July 30, 2002).

Vapnik, V. 1995. *Statistical Learning Theory*. Springer Verlag, New York.

Yu, H., J. Yang, J. Han. 2003. Classifying large data sets using SVMs with hierarchical clusters. *Proc. Ninth ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining*, ACM, New York, 306–315.