

Large Scale Detection of Irregularities in Accounting Data

Stephen Bay, Krishna Kumaraswamy, Markus G. Anderle, Rohit Kumar, David M. Steier
Center for Advanced Research, PricewaterhouseCoopers LLP
10 Almaden Blvd, Suite 1600, San Jose, CA 95113
{firstname.initial.lastname}@us.pwc.com

Abstract

In recent years, there have been several large accounting frauds where a company's financial results have been intentionally misrepresented by billions of dollars. In response, regulatory bodies have mandated that auditors perform analytics on detailed financial data with the intent of discovering such misstatements. For a large auditing firm, this may mean analyzing millions of records from thousands of clients. This paper proposes techniques for automatic analysis of company general ledgers on such a large scale, identifying irregularities – which may indicate fraud or just honest errors – for additional review by auditors. These techniques have been implemented in a prototype system, called Sherlock, which combines aspects of both outlier detection and classification. In developing Sherlock, we faced three major challenges: developing an efficient process for obtaining data from many heterogeneous sources, training classifiers with only positive and unlabeled examples, and presenting information to auditors in an easily interpretable manner. In this paper, we describe how we addressed these challenges over the past two years and report on experiments evaluating Sherlock.

1. Introduction

News headlines of recent years have contained increasingly frequent reports of corporate fraudulent financial reporting. Such frauds, which are often perpetrated by senior management, may involve very large amounts of money. In 2002, the SEC charged Adelphia Communications with fraudulently excluding \$2.3 billion of debt from its financial statements, and in that same year charged WorldCom, Inc. with overstating income by about \$9 billion. The full extent of damages from these frauds and at other companies such as Enron may never be known, but Adelphia, WorldCom, and Enron eventually filed for bankruptcy, leading to thousands of jobs being destroyed and billions of dollars in investor and creditor losses.

In response, the American Institute of Certified Public Accountants issued Statement of Accounting Standards (SAS) No. 99, which made more explicit certain expectations on auditors with respect to fraud. For example, auditors are required to apply quantitative

analytic procedures to the financial data, and to consider whether the results of those procedures identify risks of material misstatements due to fraud. Given the large amounts of data involved, and the fact that computers are at the heart of nearly all corporate financial systems, fraud risk detection is a natural candidate for software-based automated assistance.

The prior work on computer-assisted audit techniques (CAATs) for fraud detection [6] generally falls into two categories. In the first category, most of the well known analytical tests for detecting fraud risks [24] have applied ratio analysis to the consolidated financial statements, which seek to report the financial performance of a company for a year or a quarter. For example they might calculate the ratio of yearly net sales to receivables because an increase of this ratio may be indicative of premature revenue recognition, a form of financial statement fraud. However, very few actual frauds have been detected (except in hindsight) through application of such analytics. We believe that the perpetrators of large frauds take care to falsify financial statements to evade detection at such an aggregate level. The text of SAS 99 itself acknowledges "... because such analytical procedures generally use data aggregated at a high level, the results of those analytical procedures provide only a broad initial indication about whether a material misstatement of the financial statements may exist." ([1] para. 30).

In the second category, analytics and associated procedures may be applied to detailed financial data to identify individual transactions at risk of being fraudulent. Scanning analytics may be applied to identify outliers, or ratio and trend analysis normally applied to aggregated data in financial statements may also be applied to financial data that has been disaggregated, for example by business units, geographies, or time periods. While potentially more accurate than aggregate-level analytics, disaggregated analytics are much more resource-intensive, both in terms of data acquisition and interpretation of results. For example a common test is to screen for unusually large number of "round dollar amounts" (\$5000 instead of \$4893) appearing as sums of other numbers. It is not uncommon to see disaggregated analytics yield thousands of candidate transactions that then must be manually filtered further to yield a manageable set of transactions for further investigation.

We believe that effective fraud risk detection analytics should operate at the level of detailed financial transactions. Not only are these potentially more accurate, but an important side benefit is the detection of transactions that may not be fraudulent but indicate errors or control deficiencies that are also of interest to auditors. The research and development challenge is the creation of disaggregated analytics that are cost-effective to apply on a large scale. By "large scale," we mean they are suitable for use by a firm such as PricewaterhouseCoopers LLP (with over 130,000 partners and staff worldwide) which audits thousands of clients annually. Such large-scale disaggregated analytics would need to work on large amounts of data (anywhere from 1 to 50 gigabytes per client per year) and require minimum amount of data cleansing and normalization effort.

This paper describes our efforts over the past two years in addressing these goals in the context of Project Sherlock. Our primary objective is to apply analytics at a disaggregated level to find suspicious accounting behaviors that may be indicative of fraud or other material errors. We have developed a two-stage approach, where we first search for unusual behaviors in the data. These behaviors then form the features of a classifier that attempts to quantify the patterns that have been found in companies with known misrepresentations. At a disaggregated level, defining the behaviors that are interesting to an auditor is extremely difficult and the classifier effectively acts as a filter to determine which behaviors should be shown.

During the development process we faced a number of significant challenges both in the design of the statistical inference algorithms and in terms of the knowledge discovery process. From a statistical perspective, the major challenge was dealing with classification in a semi-supervised framework as the training data is composed of a small set of positive and a larger set of unlabeled examples. Specifically, the positive set was defined by financial data from companies with known fraudulent misrepresentations, and the unlabeled data consisted of companies where no misrepresentations have been discovered. It is possible to assume the unlabeled sets do not contain significant fraud or error and to treat those data sets as negative examples (as we tried at first). However in such data sets with millions of records, there is no independently reliable test that can verify the *absence* of fraud or error. We believe that a semi-supervised approach is more defensible in tolerating noisy labels and preventing bias, although care must be taken to avoid overfitting if there are too few positive examples.

From the perspective of the knowledge discovery process, we faced two additional challenges. First, obtaining financial data from multiple sources and pre-processing it into a common data model was a major effort because of the heterogeneity and scale of financial data. Without a well-defined process and automation tools

it would not be possible to even obtain a small data set for learning. Second, we needed to develop a system for explaining the results of our system to auditors who do not have a statistical or data mining background. The explanations needed to be actionable and guide auditors directly to specific accounting findings that need additional review.

In the remainder of this paper we will describe our system and how we addressed these three challenges. In the next section, we describe financial statements, the accounting process and the general ledger and related work on fraud detection. Section 3 discusses in depth the modeling approach including our exploration of semi-supervised learning and the strategy for providing detailed feedback. In Section 4, we discuss evaluation measures when there is unlabeled data and our experimental results. In Section 5, we discuss the challenges in data acquisition. Finally, we conclude the paper with a discussion of the limitations and future work.

2. The Problem of Detecting Irregularities in Accounting Data

In this section, we very briefly define and/or summarize financial statements, accounting and auditing processes, and related concepts as necessary to understand the application domain of Sherlock and prior work. This should not be taken as a thorough description of such complex subjects, for which the reader is referred to standard textbooks ([12], [21]), or better yet, consultation with a Certified Public Accountant.

2.1 Financial Statements

In the United States, the Securities and Exchange Commission (SEC) requires public companies to file reports including financial statements, typically at quarterly and annual intervals. Privately owned companies also often furnish financial statements to investors. The financial statements record the state of a company with three tables (and associated footnotes): balance sheet, income statement and cash flow statement. The balance sheet records assets, such as inventories of goods, buildings and equipment, and liabilities, such as money owed to suppliers. The income statement (also known as the profit and loss statement or statement of operations) primarily reports the earnings (profit or "bottom line") to investors and details the company's operations. The statement of cash flows completes the picture: as debt is issued from financing, assets acquired for investing and cash from sales and other operations flows into accounts, the statement of cash flows describes how these funds pay for business activities, investments and operations.

2.2 The Accounting Process and the General Ledger

The financial statements that appear in annual and quarterly reports are the highly summarized result of a complex accounting process. While the specifics of the process differ from company to company, the overall flow is shown in Figure 1. The operation of the business generates transactions that are recorded in various documents (a sales order, an invoice, a receipt, etc.). The financial impacts of these transactions are recorded in journal entries to various accounts. The cumulative impact of these entries is recorded in the general ledger (the record of activity over all accounts), and additional adjustments and consolidations are recorded in accordance with GAAP (Generally Accepted Accounting Principles), finally resulting in the financial statements reported to investors and regulators.

The primary data repository for the purposes of this paper is the general ledger (G/L). Each measurable business event results in posting a transaction in the general ledger. The transactions in the general ledger are recorded using double-entry book-keeping. Each transaction is the exchange between two or more accounts. The accounts in the general ledger describe different aspects of the business in monetary terms. Each transaction in double-entry book-keeping has a dual effect - for example, buying machinery means losing cash but gaining the monetary value of the machinery. For the accounts to remain in balance, a change in one account must be matched with a change in another account. These changes are called debits and credits. When an account is debited, another related account is credited for the same amount. The accounts to which entries can be posted - there may be thousands in a large company - are listed and numbered in a chart of accounts. The number of ledger entries in a year is often in the millions, and reaches tens or even hundreds of millions for large companies.

An accounting system will also have a number of subsidiary ledgers (called subledgers) for items such as cash, accounts receivable and accounts payable. All entries that are entered (posted) to these subledgers will, at least in summary form, transact through the general ledger accounts. There are also times when transactions are posted directly to the general ledger, including capital financial transactions or adjustments.

Entries are posted to accounts by updating account balances to reflect credits and debits noted in the entry. The general ledger accumulates all transaction information in one place for all the accounts of a company. A trial balance lists the ledger accounts and their balances at a particular point in time. The financial statements (consisting of the income statement, balance sheet, and statement of cash flows) are derived from the trial balances and the chart of accounts. The contents of a

general ledger vary from company to company but at least have the date of the transaction, the accounts that are debited or credited and the amounts involved in the transaction.

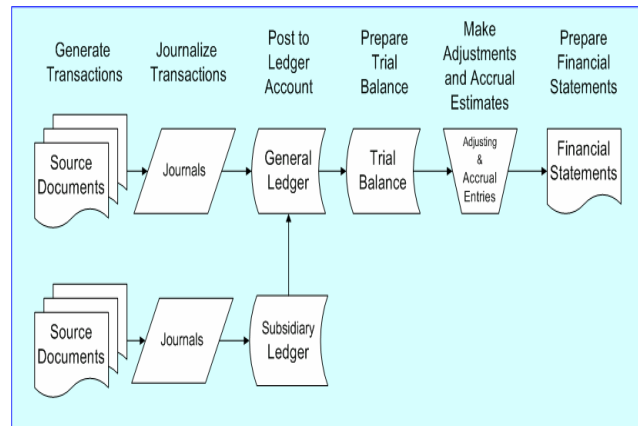


Figure 1: Overview of the accounting process

2.3 Sherlock's role in detecting irregularities in accounting data

In general, the task of the auditor is to express an opinion on whether the financial statements of a client are free of material misstatement (for our purposes, the word "material" can be informally interpreted as "significant"). The target of our research on Sherlock is to assist the auditor in detecting irregularities in the financial accounting data contained in the general ledger that might represent misstatements. The irregularities fall into three categories:

- Irregularities due to financial statement fraud. Two common classes of financial statement fraud include improper revenue recognition and asset misappropriation.
- Unintentional errors, such as clerical mistakes when entering data that have a material effect on the financial statements.
- Unusual entries in the general ledger that an auditor would consider worth investigating, even if they are ultimately found to be neither fraud nor error.

These three categories are required because it is not possible to determine the presence of fraud from the data alone: a determination of fraud requires the involvement of forensic specialists and a legal finding of intent to commit fraud. Furthermore, auditors are interested in understanding unusual situations and aggressive accounting practices, because if management's interpretations of accounting rules are questioned by

investors and/or regulators, those are most likely to lead to restatements, with potentially severe consequences for the company and their auditors.

On the other hand, flagging too many irregularities that are false positives because they turn out not to fall into the three categories above also has negative consequences. We need to take care that deviations from the norm that occur during the normal operations of a business, such as seasonal variations, do not in themselves cause an irregularity to be flagged for follow-up. Auditors operate under tremendous time pressure, since they must perform the majority of their final work and render an opinion in the short time between the time a client closes its books and the deadline for filing with the SEC. Investigation of irregularities may involve interviews, requests for supporting documentation and demand for client time that is also in short supply. Since the true incidence of irregularities that might cause material misstatements is probably quite low (we hope), it is critical for Sherlock's acceptance to minimize false positives, which are more likely than false negatives in this application.

A final disclaimer before proceeding to discuss related work and the body of the paper: Sherlock is research in progress. As such, the methods we describe should not be interpreted as descriptive of PwC's current standard practice in analyzing general ledger data.

2.4 Related work on detecting irregularities

Auditing standards require that an audit targets high risk areas and mitigate those risks using substantive testing. Statistical sampling [15] techniques have been employed to test the accounting. The sampling techniques help auditors quantify the extent to which their procedures have mitigated those risks. Also used are "*red flags*" that indicate irregularities compared to peer companies and to historic values.

In addition to the sampling-based approaches to detect accounting irregularities, there have been many model-based approaches developed to help find these irregularities. Trend, ratio, and regression analysis have all been used to identify unexpected and unusual activity that indicates higher risk of material misstatement. Most of the work in the area of detecting accounting irregularities are developed using data that is aggregated at some level, such as financial statements, business unit, or geographic territory. Statistical methods for finding accounting irregularities usually use financial statement ratios to find unusual changes to the ratios over time. Generalized linear models have been used on financial ratios to predict companies that are cited for financial statement misrepresentation by the SEC ([2], [3], [14]). Typically, a logistic regression model is used to predict likelihood of fraudulent financial reporting based on the presence or absence of fraud risk factors. These risk factors include several financial statement ratios and also

include indicators like rapid company growth, inadequate or inconsistent relative profitability, etc.

A generalization of the above classification models using more flexible link functions led to the use of neural networks for fraud detection ([13], [17], [19]). Artificial neural networks have also been used to learn models for detecting management fraud using publicly available predictors of fraudulent financial reporting [9]. The correlation of events related to irregular accounting practices have been explored in studies that investigate short selling, insider trading, CEO compensation, likelihood of bankruptcy and restatement [8].

Most of these models use labeled data with verified accounting irregularities for developing a classification model that can then be applied for prediction in future data sets. When a financial statement is classified as having high risk or fraud during the preliminary stage of an audit, it signals the auditor to increase substantive testing during the fieldwork.

Among other related work, Bolton and Hand [4] discuss some of the techniques for fraud detection in areas such as money laundering, e-commerce, credit card fraud, telecommunications fraud, computer intrusion, etc. Chan et al. [5] also discuss some of the data mining approaches for detecting credit card frauds. These approaches target specific domains and shows an increase in the use of statistical and data mining techniques for fraud detection.

3. Modeling approach

Our goal is to find irregularities in general ledger data that an auditor would find suspicious enough to follow-up. However, "suspicious" is difficult for auditors to specify exactly or quantify and hence we choose to operationalize our goal as discovering accounting behaviors that are

- unusual both with respect to a company's past history and to its peers, and
- identified by a classifier as being similar to those in companies with known misrepresentations.

We believe this combination is necessary because searching solely for anomalies in a general ledger is too unguided. In our experience, pure anomaly detection algorithms yield far too many results that are not of interest to auditors because of two reasons: (1) many times a company will record business transactions in a completely different fashion from other companies for perfectly legitimate reasons, and (2) companies have many normal events that occur only once within the scope of data that was collected and hence would appear very unusual.

However, we believe that looking for anomalies is necessary because although auditors have general ideas about the types of behaviors that are suspicious, their

evaluation depends on the magnitude of the effect with respect to their expectations, which in turn depends on the company's past history and the activity of peers. For example, transactions representing the return of sold items is only suspicious if it occurs in a much larger volume than expected based on the past history of the company and the return rate of competitors selling similar products.

The classifier is necessary for two reasons: First, a classification algorithm will automatically learn how to quantify the effects and can determine their relative weightings. Second, a classifier provides an overall assessment of the general ledger that helps in the allocation of resources available for follow-up.

In Sherlock, our modeling approach is based on three stages:

1. Develop features based on irregularities or other unusual behaviors in the general ledger. The features track characteristics known to be useful by auditors and investigators.
2. Train a classifier using the features to characterize general ledgers with known manipulations versus those without.
3. Using the outputs of the classifier, assign credit to specific accounts and transactions by working backwards through the definitions of the model and features.

There are two important constraints that affect the choice of a classifier in stage 2. First, although we have data from several different general ledgers for training statistical models, the state of the G/L with respect to the presence of fraud or error is known only asymmetrically. Specifically, there are two types of data available:

1. General ledger data from companies with known financial statement misrepresentations.
2. Data from companies where no financial misrepresentation has been discovered to date.

Clearly, data from the first source will contain positive examples of the phenomena that we are trying to capture with our statistical models. However, data from the second source is problematic from a modeling perspective since it may be inaccurate to simply treat them as negative examples that are free from misrepresentations. There may exist misrepresentations that have not yet been discovered. Furthermore, our goal also includes the detection of suspicious irregularities which may be present even if there is no fraud or error.

A second important constraint is the unavailability of costs associated with false positive and negative errors. Traditionally, in most detection systems, the designer adjusts the parameters to minimize the total expected cost while recognizing that false positives and negatives may have different consequences. In our domain, however, precise cost information is currently unavailable and

difficult to obtain. For false positives, the cost is determined by the time required by auditors to follow-up on a finding and this can vary widely from just a few minutes if there is a well known business explanation, to many hours if it involves examining supporting documentation. With enough trials, it might be possible to obtain a distribution for these costs. A larger problem is quantifying the cost for false negatives, which could vary from zero to millions of dollars if a fraud or error is later discovered and results in litigation. However, litigation is a rare event and it would be difficult to accurately quantify this risk even with a long observation period because of the extreme outcomes. Furthermore, if the false positive rate is too high, the system will be seen as unreliable by the auditors and they may not treat the findings seriously even if the system is operating at a false positive / negative ratio that minimizes the expected cost.

Finally, we note that we analyze each general ledger by quarter, corresponding to the most frequent periodic reporting typically required by the SEC.

3.1 Engineering features that flag irregularities

In this section, we describe how each quarter of a general ledger can be represented by a feature vector where each element denotes whether or not there was irregular activity.

We chose to represent each G/L as a feature vector for two reasons: First, general ledgers can be extremely heterogeneous. Depending on the company, a G/L might contain anywhere from few hundred accounts to as many as twenty-five thousand and there exists no standard mapping from the accounts in one company to those in another. Furthermore, depending on the company's financial system and how business transactions are recorded into the general ledger from various sub-ledgers, the number of transactions can vary from a few hundred to millions. Developing features allows us to represent general ledgers as points in a standard vector space as is typically assumed by most learning algorithms. Second, representing the general ledger as a feature vector allows one to abstract useful characteristics for learning algorithms while ignoring irrelevant information.

In general, our design philosophy has been to develop features that measure changes in behaviors that might be interesting and suspicious from the viewpoint of the auditor. We developed over fifty features that quantify the activity within a general ledger according to its magnitude, timing, interactions with other activities, and associated descriptive properties.

For each feature, we first compute a raw score which can be thought of as a number representing the absolute strength of the activity for the company. This number is then normalized by forming a ratio to a measure of historical activity that is based on the raw scores for the previous four quarters.

Having features that are ratios allows numbers to be compared across companies as they represent dimensionless growth numbers. Forming the ratios also reduces statistical dependencies between quarters in a similar fashion to differencing. For example, if we had a feature that looked at dollars in transactions of a specific type, we obviously would expect higher values to occur in companies that are larger. Normalizing by forming a ratio mostly eliminates the effect of company size.¹

After the features have been normalized, we discretize the continuous values to 0 or 1 by comparing the ratio to a threshold set such that only the top 5% of data points from companies with no discovered misrepresentations will be assigned a value of 1.

We chose this discretization approach because we believe that only extreme behaviors are informative for our prediction task. Setting the discretization boundary at a 5% rate ensures that a feature only takes a value of 1 when the underlying behavior is more unusual than most other companies. This discretization approach also has the benefit that it is easy to explain to the auditors (i.e., exceeding a discretization threshold can be viewed as setting off a risk indicator).

3.2 Classification with positive and unlabeled training data

The features will typically highlight irregular activity that is beyond the normal range of behavior for a company. However, as discussed in a previous section, there could be many behaviors in a company that appear anomalous or unusual for perfectly legitimate reasons. Thus we use a classifier to identify those behaviors that should be given extra review by the auditors. The classifier is also used to provide a global assessment of how suspicious a company might appear and this assessment is used for prioritizing which general ledgers receive extra scrutiny.

The main issue when developing and choosing a classifier for this stage was dealing with positive and unlabeled data. As discussed previously, we have data from companies with known misrepresentations that can be considered positive examples. However, we also have data from companies where no misrepresentations have been discovered but which might have contained suspicious behaviors (that ultimately turned out to be legitimate, but worthwhile to investigate). Data from these latter companies are considered unlabeled.

We considered and experimented with many different classification algorithms, but in this paper, we focus our

discussion on two extensions of the naïve Bayes classifier that we felt were the most promising:

- Positive naïve Bayes ([7], [18]) estimates probabilities for a two class binary problem explicitly assuming the presence of a positive set of examples and an unlabeled set.
- Naïve Bayes using the EM algorithm for estimating parameters when the class variable may be hidden for specific training points.

In the naïve Bayes classifier, dealing with positive and unlabeled data causes an estimation problem because the normal method of estimating the class conditional probabilities is based on counting the frequencies of observed features and dividing by the number of class examples. Since there are no explicit negative examples, and therefore no counts, we cannot apply a standard implementation of naïve Bayes. Instead of the standard procedure, the Positive Naïve Bayes (PNB) algorithm proposed by Denis et al. [7] modifies the training procedure in order to handle unlabeled data. Specifically, PNB estimates the conditional probabilities for the negative class by using the mixture component assumption

$$P(x) = P(x|y=1)P(y=1) + P(x|y=0)P(y=0) \quad (1)$$

where $y=0$ is the negative class and $y=1$ is the positive class. Since $P(x)$ and $P(x|y=1)$ are known or can be estimated from the data, the probabilities for $P(x|y=0)$ can be computed by solving the linear equation assuming the priors, $P(y=0)$ and $P(y=1)$, are also known.

Other work on the problem of learning from positive and unlabeled data has addressed the challenge by estimating $P(X|y=0)$ indirectly from the identification of likely negative samples. Lin et al. [19] employ a spying technique (S-EM) that uses known positive samples in the unlabeled set (spies) to calibrate the retrieval of likely negative samples. A successive EM step is used on all data (labeled and unlabeled) for building a classifier, utilizing the likely negative points to initialize the algorithm. This approach can be understood as a supervised initialization of the EM algorithm, and be regarded as an extension to initializations that considers all unlabeled data negative.

The transformation of continuous values into binary features results in a parametric naïve Bayes formulation, where the class conditional probabilities become multivariate Bernoulli, thus the generating model becomes a mixture of multivariate Bernoulli distributions. This model has been used in text classification [16], where the unlabeled data in form of missing data is incorporated via the EM algorithm, as it is not uncommon to have large amounts of text documents without labels.

¹ Note that the ratio does not completely eliminate all effects of company size because the variance can be larger for smaller companies.

The EM algorithm may converge to parameter estimates that lead to poor classifier performance due to violations of the equivalence of mixture components and classes. Remedies include the introduction of constraints such as a class distribution constraint by calibrating the probabilistic posteriors to maintain fixed proportions of negatives and positives [23]. It was also proposed to restrict the full update of the EM algorithm usually applied to all parameters (class priors, posteriors) to only unlabeled data while keeping labeled data parameters fixed, therefore separating the unsupervised and supervised learning tasks [11].

3.3 Providing actionable feedback

The classification model produces an overall score for each quarter of a general ledger and this can be interpreted as a measure of how suspicious it is as a whole. However, this knowledge is insufficient for auditors who need feedback at a more detailed level to perform follow-up procedures. As part of their regular duties, auditors often test specific accounts and transactions for correctness and hence we would like to provide feedback at this level.

We treat the problem of identifying accounts and transactions as one of credit assignment where we assign credit (or blame) to individual elements of a G/L for the high overall score. Our basic approach is to work backwards through the equations defining the naïve Bayes model and the features to identify causal accounts and transactions. Specifically, we use the following three step process:

1. Determine the features in the naïve Bayes model responsible for a high overall score.
2. For each feature identified in step 1, find the accounts that contribute most to the feature's normalized value.
3. For each combination of feature and account, determine the transactions responsible for the high contribution of the account to the feature's value and select those that are most likely to be of interest to the auditors.

Fortunately, linear models make it easy to step backwards and identify the inputs causing high scores. In step 1, we find the features in the naïve Bayes model that have a positive weight of evidence, defined as follows:

$$w_i = \log P(x_i | y = 1) - \log P(x_i | y = 0) \quad (2)$$

We identify all features where w_i is positive and $x_i = 1$ as contributing to the high score. Note that we require $x_i = 1$ because this corresponds to an unusually large score with respect to the past history of the company (see Section 3.1). In general, $x_i = 0$ is not informative to the

class decision and we have never observed w_i substantially different from 0 when $x_i = 0$.

Having identified the features, step 2 involves identifying the key accounts causing the high score for that feature. This clearly depends on each feature's definition, but in general, the features are often computed by aggregating scores from individual accounts to form a final value. Thus, we can examine the contribution of each account and pick those that have the largest effect.

With the feature and accounts identified, we then proceed to step 3. We first extract all transactions that contribute any non-zero amount to the feature's value. Usually, this will result in many transactions that are far too numerous for the auditors to evaluate. Thus we limit these transactions by selecting a small subset according to the following two criteria:

- Select transactions with the largest magnitude, i.e., we would prefer a transaction representing a \$5,000,000 debit over one for say \$250,000.
- Select the transactions that appear most unusual compared to the previous quarter. Every transaction can be thought of as representing a flow of money where a source account is credited and a destination account is debited. We define unusualness by examining the pair of accounts being debited/credited and prefer those with the largest difference in flow compared to the previous quarter.

To summarize, we have taken a credit assignment approach that begins with an aggregate score and from that determines specific elements of the general ledger to flag for additional review.

4. Evaluation of models

In this section, we evaluated Sherlock on its ability to flag general ledger quarters with suspicious behaviors. A major challenge here is determining performance in the context of unlabeled data where a prediction cannot be directly validated as being right or wrong and thus traditional measures such as recall and precision cannot be computed.

Instead we will use several alternate measures that attempt to quantify, as much as possible, the same ideas of sensitivity and selectivity. Specifically we use the following measures:

- **Recall rate on known fraudulent quarters.** This is a measure of sensitivity, the ability to detect items that should be flagged. If the fraud and unlabeled points are drawn from the same joint distribution, then this measure may be a good surrogate for the overall recall rate.

- **Positive rate on unlabeled data.** This is a coarse measure of selectivity. In general, we expect that the optimal classifier would flag a small, but non-zero, number of unlabeled data points.
- **Pessimistic estimate of lift.** Lift charts display how well a classifier is performing relative to a random selection as the number of candidates marked as positive increases. The pessimistic estimate assumes that all unlabeled instances are true negatives.
- **Pessimistic estimate of the ROC curve.** The ROC curve [10] explicitly characterizes the tradeoff between the false positive and true positive rates for a classifier as a decision threshold is varied. The pessimistic estimate assumes that all unlabeled instances are true negatives and any such points that are flagged by a classifier are false positives.
- **Number of negative weights of evidence.** The sign of the weight of evidence, as defined by Equation 2, can be thought of in a similar fashion to the sign of a regression coefficient. For our features, a positive weight of evidence matches our belief that extreme behavior is more closely linked to the known fraudulent companies. As in linear regression, models where the signs do not conform to expectations may be an indication of a poor model and they hinder acceptance and understanding by domain experts [22].

Note that in this paper, we did not evaluate our methods for providing actionable feedback (Section 3.3), because the available data did not denote in the positive data sets the specific accounts or transactions that were deemed worthy of additional investigation.

4.1 Experimental procedure

We evaluated naïve Bayes, PNB, and the EM approach discussed in the previous sections. In addition, we also compared the results to logistic regression. We used the following experimental setup:

- The positive data consisted of 7 known fraudulent companies with 67 quarters and the unlabeled data was formed from 35 companies with 70 quarters. The selection of the unlabeled set was deliberately biased towards higher risk companies.
- The naïve Bayes and logistic regression classifiers were trained assuming that all unlabeled points are negative.
- Probabilities corresponding to zero frequency counts were smoothed with virtual points. For naïve Bayes, PNB, and EM we set the number of points to 0.5.
- The prior for training naïve Bayes and PNB was set to 0.2. Although published literature reports that the incidence of financial misrepresentation is small ([3]; [14]), recall that we are trying to find suspicious activity which may occur much more broadly.
- For logistic regression, we developed models using both the continuous and discretized features. However, the performance was very poor with continuous data and hence we only report results on the discretized features.
- All classifiers were trained with points weighted such that each company year was given equal weight. For example, if a company had two consecutive quarters, they would each be given a weight of 0.5.
- We used a modified cross-validation approach where all quarters from a company are assigned to an individual fold. This prevents positive bias in performance estimates compared to assigning quarters randomly, because quarters from the same company cannot be in both the training and test groups. Discretization was performed inside the cross-validation process.

4.2 Performance on statistical measures

The results for recall on known fraudulent points and the positive rate on unlabeled data are shown in Table 1. At the native threshold for the classifiers, PNB achieves the highest recall rate on known fraudulent cases. However, PNB also has a high rate of flagging positives on the unlabeled data. EM and NB score similarly on both measures although EM has slightly better recall on known frauds and flags fewer unlabeled cases. Logistic regression achieved a high recall rate but also obtained a high positive rate on unlabeled data points. Considering that the number of positive and unlabeled data points is similar, logistic regression is only a little better than random.

Table 1 also shows the number of positive weight violations. PNB had the fewest sign violations, followed by naïve Bayes and EM. Logistic regression had the most violations. As discussed earlier, negative weights of evidence are problematic for domain experts since they imply that more extreme behavior in the G/L decreases risk.

Table 1: Recall, Positive Rates, and Sign Violations

	PNB	NB	EM	LR
Recall Known Frauds	64%	34%	37%	55%
Positive Unlabeled	27%	10%	8%	44%
Sign Violations	3	8	10	21

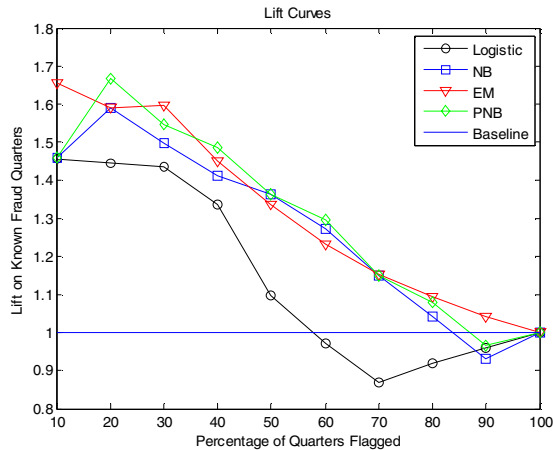


Figure 2: Pessimistic Lift Curves for PNB, NB, EM and Logistic Regression

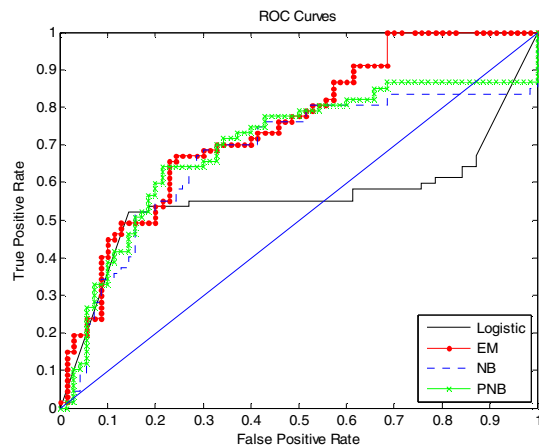


Figure 3: Pessimistic ROC curves for PNB, NB, EM and Logistic Regression

Figure 2 shows a pessimistic lift curve for the four classifiers. The pessimistic curve assumes that all flagged unlabeled points are false positives and hence the true performance may be greater. On the left hand side of the graph, PNB and EM have the best performance, although NB performs surprisingly well and closely follows PNB. Logistic regression is the worst performer of the classifiers. We are primarily concerned with performance on the left hand side of the lift curve in order to maximize the ratio of true to false positives under the resource constraint that only a limited number of cases can be marked for additional follow-up. Both PNB and NB have a dip in the lift curve at 10%. The dip is caused by both classifiers flagging several unlabeled points as positives with high scores.

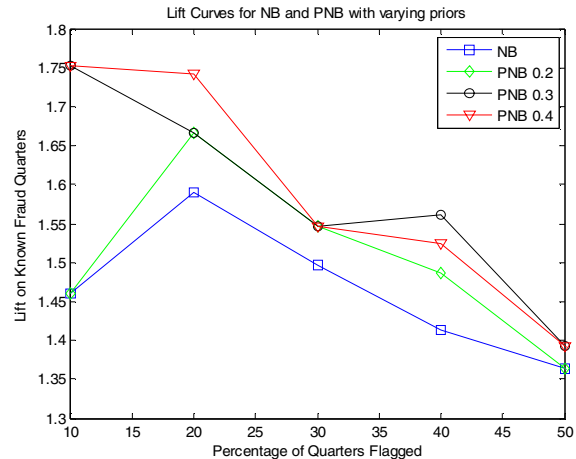


Figure 4: Lift curves for NB and PNB with varying priors

Figure 3 shows pessimistic ROC curves for the four classifiers. For a false positive rate of 0% to 15%, the curves are similar, but beyond this point, logistic regression deteriorates rapidly. As with the lift curve, NB is close, but slightly worse than PNB and EM. Note that the ROC curve is normally a step function, but for logistic regression we see diagonal line segments at both the beginning and the end of the curve. This occurs because it classifies many points as positive with probability one and negative with probability zero (to machine precision).

The results from figures 2 and 3 were surprising to us as we had expected larger differences between PNB and EM, which both make the assumption of positive and unlabeled data points, compared with naïve Bayes, which makes the assumption that the labels are completely correct. As a result, we hypothesized that perhaps the classifiers performed similarly because of an inappropriate choice of prior which we had set to 0.2 in the initial experiment. Thus we ran another experiment varying the prior from 0.1 to 0.5. We focused the experiment on PNB and NB since these are closely paired classifiers whose only difference is in the estimation of $P(x | y = 0)$.

In Figure 4 we show the results for NB and PNB with varying priors. The results for PNB with prior = 0.1 and prior = 0.5 are not shown to reduce clutter and respectively, they were similar to the results for NB and PNB with prior = 0.4. Note that the lift curve for naïve Bayes does not change with the prior as differences between $P(y=0)$ and $P(y=1)$ manifests itself as a constant factor in the computation of $P(y|x)$ and is identical for all cases. At low values of the prior, PNB and NB perform similarly, but as the prior increases so does PNB's advantage over NB. At priors of 0.3 and 0.4, the PNB has a large lift advantage (30%) over NB when the first decile of quarters are flagged. These results suggest that the

unlabeled assumption can substantially increase power when the prior is set to an appropriate level.

In summary, we note that in our initial experiments PNB and EM performed similarly to NB which does not account for unlabeled data and instead assumes that the points are true negatives. However, further investigation varying the prior on PNB revealed that for higher settings the performance gap in favor of PNB increased substantially. Logistic regression had a competitive ROC curve at low false positive rates but the high number of sign violations would hinder its acceptance and make explanations difficult.

Finally, we note that there are several limitations to our analysis here which should be considered when interpreting the results. Specifically, our leave-one-company-out approach can yield pessimistic estimates since there are very few positive companies. Leaving even a single company out of the training can have a large effect. Additionally, our data points are not independent since they are related by company association. While we have avoided major bias by using leave-one-company-out cross-validation, having dependent quarters can still inflate the variance of our performance measurements.

5. Data Challenges

Before concluding, we address a practical issue that needed to be resolved before any of the previously described analytical work could be performed. The general ledger data for different companies may be maintained by homegrown software or an Enterprise Resource Planning (ERP) package from any of dozens of vendors. Each system has its unique way of storing and maintaining the financial information. Even for companies using the same ERP system, there is a wide variety in the customization of the systems. Each company also tracks varying levels of details of their financials using these ERP systems. In the long term, this diversity may be reduced, both as the ERP software market consolidates, and as standards such as XBRL are adopted to facilitate the interchange of data. In the short term, however, those wishing to use analytics on G/L data must be prepared to encounter difficulties in applying the same analytics to detailed accounting data from different companies.

The current practice for analyzing G/L data is to write PL/SQL code or custom scripts using a tool such as ACL to extract the data in a system-specific manner². Previously the range of systems was such that there was no alternative to custom extraction. Increasingly, as companies begin to adopt one of a small set of ERP software packages for maintaining their financial records, there is some benefit to be gained for automation for those

packages. The data management system behind Sherlock uses a variety of techniques based primarily on a common data model for general ledger data that PwC has constructed using the XBRL GL standard taxonomy³ as a foundation. A commercial ETL (Extract, Transform and Load) tool is used to map, transform and load the clients' source data files into this model. ERP-specific mapping rules are used in mapping the source data to the G/L Common model. The mappings are also customized based on the clients' customization of their ERP system.

Based on our experience, we believe we have reduced the time to extract transform, load, and validate the data from a G/L by an order of magnitude from earlier manual processing.

6. Limitations and future work

Our work on Sherlock represents the beginning of our efforts to analyze detailed accounting data. While we are pleased with our initial results, our experience working with the data suggests that certain areas can be improved. In this section, we discuss several limitations of our approach and our plans for addressing them.

From a design perspective, our method is based on detecting unusual changes in a company's G/L from its past history. We focused on changes because it is a very effective way of controlling for company size and behaviors that are specific to the company but not illegitimate. However, a change based-system will not catch activities that only represent gradual changes or activities that were initiated before the timeframe covered by the data. At present, we believe focusing on changes is critical to keeping the false positive rates low, but one possible direction around this problem is to use alternate normalization criteria such as a measure of the company size in the current time period or by ranking the company according to close peers.

In our statistical models, we have an important limitation as the quarters of general ledger data that compose our training sets are clearly dependent and this may lead to biased estimates of the naïve Bayes coefficients. We expect to address this issue in the future by extending the naïve Bayes model to incorporate links between the class node in consecutive time points (similar to hidden Markov models with a hidden state for each quarter but with multiple observable outputs).

We also plan to improve our process for extracting accounts and transactions. Currently, it is treated as a credit assignment problem and does not incorporate any learning or direct feedback from auditors as to the accounts or transactions that may be interesting. We hope to obtain enough feedback at the level of accounts and

² http://www.acl.com/pdfs/ACL_V8_Fact_Sheet.pdf

³ <http://www.xbrl.org/GLTaxonomy/>

transactions to consider supervised and semi-supervised approaches.

7. Conclusions

We presented a system for identifying suspicious irregularities in detailed financial data, specifically the general ledger of a company. Our method is based on developing features that catch irregularities in the data and then applying a classifier to identify those irregularities that occur more commonly in known suspicious entities.

From a statistical perspective, the primary challenge we faced was dealing with the issue of unlabeled data in our classifiers. That is, for many of the training data points the class label is not known with absolute certainty and they may in fact be either positives or negatives. To address this, we experimented with two variations of naïve Bayes that consider the effect of unlabeled data in parameter estimates. The results indicate that explicit treatment of label uncertainty can improve performance given appropriate assumptions.

We also discussed how we addressed two other important challenges: reporting information to domain experts in an actionable and understandable manner, and finally, organizing the data collection process to be efficient on a large scale.

Acknowledgements

This work, begun in August 2003, was conducted within the PricewaterhouseCoopers' Center for Advanced Research (CAR), directed by Sheldon Laube and Glenn Ricart. In addition to their support and feedback, we'd like to acknowledge the assistance of our colleagues within CAR: Jeff DeLisio, Mave Houston, Li Chen, Eric Berg, Lever Wang and several PwC staff who worked with us on temporary assignment to CAR: Luis Furtado, Nathan Kendig, Erick Vargas, Yan Tong, Sunny Tsoi, Mehul Patel, Anna Watson, and Leena Mansharamani. Student interns and other consultants included Jimeng Sun, Madhuri Shah, Xue Bai, Tim Anspaugh, and H. T. Hu. People too numerous to list here, both from the Assurance practice of PwC and from various support functions within the firm, provided assistance, but several leaders and groups deserve special mention: Pat McNamee and the staff of the Global Audit Methodologies Group who helped us understand the auditing process, Robin Cutshaw and others in the Data Management Group who helped us get the data, and Philip Upton and other members of the Forensic Technologies Solutions group who contributed key insights, data and support. Finally, we are grateful for comments on an earlier draft of this paper from Pat McNamee, Galit Shmueli, Wolfgang Jank, Christos

Faloutsos, Santosh Ananthraman, Jessica Pollner, Hakan Gogtas, and Charis Kaskiris.

8. References

- [1] American Institute of Certified Public Accountants, Inc. (2002). Consideration of Fraud in a Financial Statement Audit, SAS No. 99, 2002.
- [2] Bell, T., & Carcello, J. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice and Theory*, 19, 169-184.
- [3] Beneish, M. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, 55(5), 24-36.
- [4] Bolton, J. & Hand, D. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, 17, 235-255.
- [5] Chan, P., Fan, W., Prodromidis, A. & Stolfo, S. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems*, 14(6), 67-74.
- [6] Coderre, D.G. (1999). *Fraud Detection: Using Data Analysis Techniques to Detect Fraud*. Vancouver, B.C.: Global Audit Publications.
- [7] Denis, F., Gilleron, R., and Tommasi, M. (2002). Text Classification from Positive and Unlabeled Examples. In *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*.
- [8] Desai, H., Krishnamurthy, S., and Venkataraman, K. (2004). Do Short Sellers Target Firms with Poor Earnings Quality? Evidence from Earnings Restatements. Available at SSRN: <http://ssrn.com/abstract=633283>.
- [9] Fanning, K., & Cogger, K. (1998). Neural Network Detection of Management Fraud using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance, & Management*, 7, 21-41.
- [10] Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27, 861-874.
- [11] Ghahramani, Z. & Jordan, M.I. (1994). Supervised learning from incomplete data via an EM approach. In Cowan, J.D., Tesauro, G., and Alspector, J. (eds.). *Advances in Neural Information Processing Systems*.
- [12] Golden, T.W., Skalak, S.L., & Clayton, M.M. (2006). *A Guide to Forensic Accounting Investigation*. Hoboken, NJ: John Wiley & Sons.

- [13] Green, B.K. & Choi, J.H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing*, 16(1), 14-28.
- [14] Grove, H. & Cook, T. (2004). A Statistical Analysis of Financial Ratio Red Flags. *Oil, Gas and Energy Quarterly*, 53(2), 321-346.
- [15] Hitzig, N. (2004). Statistical Sampling Revisited. *The CPA Journal Online*. Retrieved from <http://www.nysscpa.org/cpajournal/2004/504/essentials/p30.htm>.
- [16] Juan, A. & Vidal, E. (2002). On the use of Bernoulli mixture models for text classification. *Pattern Recognition* 35(12): 2705-2710.
- [17] Koskivaara, E. (2004). Artificial Neural Networks in Auditing: State of the Art. *The ICFAI Journal of Audit Practice*, 1(4), 12-33.
- [18] Letouzey, F., Denis, F., and Gilleron, R. (2000). Learning from positive and unlabeled examples. In *Proceedings of the Eleventh International Conference on Algorithmic Learning Theory*, 71-85.
- [19] Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. *Managerial Auditing Journal*, 18, 657-665.
- [20] Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially Supervised Classification of Text Documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*.
- [21] O'Reilly, V.M., McDonnell, P.J, Winograd, B.N, Gerson, J.S, & Jaenicke, H. R. (1998). *Montgomery's Auditing*, Twelfth Edition. New York: John Wiley & Sons.
- [22] Pazzani, M.J. and Bay, S.D. (1999) The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*.
- [23] Tsuruoka, Y. & Tsujii, J., (2003). Training a Naive Bayes Classifier via the EM Algorithm with a Class Distribution Constraint. *Proceedings of CoNLL-2003*, 127-134.
- [24] Wells, J.T. (2004). *Corporate Fraud Handbook: Prevention and Detection*. Hoboken, NJ: John Wiley & Sons.