Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue

Kuo-Wei Hsu, Nishith Pathak, Jaideep Srivastava,
Greg Tschida and Eric Bjorklund

Abstract We present a case study of a pilot project that was developed to evaluate the use of data mining in audit selection for the Minnesota Department of Revenue (DOR). The Internal Revenue Service (IRS) estimated the gap between revenue owed and revenue collected for 2001 to be approximately \$345 billion, of which they were able to recover only \$55 billion, and the estimated gap for 2006 was approximately \$450 billion, of which the IRS was able to recover only \$65 billion. It is critical for the government to reduce the gap and the fundamental process for doing so is audit selection. We present a data mining based approach that was used to improve the audit selection process at the DOR. We describe the manual audit selection process used at the time of the pilot project for Sales and Use taxes, discuss the data from various sources, address issues regarding feature selection, and explain the data mining techniques used. Results from the pilot project revealed that the data mining based approach can increase efficiency in the audit selection process. We also report results from actual field audits performed by auditors at the DOR, and results validated the usefulness of the data mining based approach for audit selection. The impact of the pilot project would be a refinement of the manual audit selection process and tax assessment procedures for other types of taxes.

K.-W. Hsu (✉)

Department of Computer Science, National Chengchi University, Taipei, Taiwan (ROC)
e-mail: hsu@cs.nccu.edu.tw

N. Pathak · J. Srivastava

Department of Computer Science and Engineering, University of Minnesota,
Minneapolis, MN, USA
e-mail: npathak@cs.umn.edu

J. Srivastava

e-mail: srivasta@cs.umn.edu

G. Tschida

Department of Revenue, State of Minnesota, St. Paul, MN, USA
e-mail: greg.tschida@state.mn.us

E. Bjorklund

Computer Sciences Corporation, Falls Church, VA, USA
e-mail: ebjorklu@csc.com

1 Introduction

The Internal Revenue Service (IRS) uses the concept of *tax gap* to estimate the amount of non-compliance with tax laws.¹ The tax gap measures the difference between the amount of tax that taxpayers should pay and the amount that taxpayers actually pay on time. The estimated tax gap is stable over time, ranging between 16 and 20 % of tax liability [29]. For Tax Year 2001, the estimated tax gap was roughly \$345 billion and only a small portion was eventually collected. The IRS recovered roughly \$55 billion, 15.9 % of the total tax gap, reducing it to \$290 billion for Tax Year 2001. For Tax Year 2006, the estimated tax gap was about \$450 billion, while the net tax gap (which could not be collected through enforcement or late payments) was about \$385 billion. Tax evasion is problem that all modern economics have to face and solve [34]. As former IRS Commissioner Mark W. Everson indicates,²

“The magnitude of this tax gap highlights the critical role of enforcement in keeping our system of tax administration healthy.”

Also as Andreoni et al. indicates [1],

“Characterizing and explaining the observed patterns of tax noncompliance, and ultimately finding ways to reduce it, are of obvious importance to nations around the world.”

Improving government efficiency is important for effective governance, while improving tax assessment efficiency is essential for economic activities. This is especially critical in a tough economy.

Since tax is the primary source of revenue for the government, it is imperative for the government to reduce the tax gap. The first step in doing so is to understand the sources of this tax gap. These include non-filing of tax returns, underreporting of tax, and underpayment of tax. Underreporting of tax is the single largest factor, and the amount underreported is much larger than the sum of non-filing and underpayment. As Toder indicates [29],

“Underreporting of tax liability is a much bigger source of the tax gap (\$285 billion) than either underpayment (\$33 billion) or non-filing (\$27 billion).”

Thus, underreporting of taxes is an important challenge presented to the government; however discovering these cases requires considerable effort and work from multiple departments. The problem for the government is how to efficiently discover individuals or businesses that potentially owed taxes [11]. The data mining community has made contributions towards solving this problem. Related work includes using artificial neural network (ANN) to determine if an audit case requires further audit [35], using classification techniques to assist in strategies for audit planning [4, 5], and using machine learning and statistical methods to identify high-income individuals taking advantage of abusive tax shelters [14].

¹ <http://www.irs.gov/newsroom/article/0,,id=158619,00.html>

² <http://www.irs.gov/newsroom/article/0,,id=154496,00.html>

Most research models underreporting of tax as fraudulent behavior and discovery of unreported taxes as a fraud detection problem. Business or finance related fraud detection problems have been receiving a lot of attention from the data mining community [3, 5, 24, 25, 31, 32]. In the pilot project presented in this article, in contrast, we focused on a different problem, viz. audit selection. An audit is an exploration into records of a taxpayer with the goal of finding out if the taxpayer properly reported tax liabilities, and the goal of an audit selection strategy is not to track down all tax evaders but to use the resources in a more efficient way in order to help the government get better returns on investments [19].

In particular, in the pilot project, we focused on a specific form of tax called Use tax. Use tax is similar to but differs from Sales tax in that Use tax is on the use of taxable goods and services [13]. However, the data mining based approach presented in this article also helped audit selection for Sales tax, since audits for Sales and Use taxes are usually conducted together. Cornia indicated that Sales tax could contribute to 30 % of the annual revenue of states that have employed Sales tax (and most states have done so), and also that the estimated loss in Use tax revenue could be more than \$55 billion by 2011 [13]. While the audit selection process for Sales and Use taxes is an important part of tax administration, so far (to the best of our knowledge) limited research in the areas of data mining applications has addressed it.

Bots and Lohman discussed the added value of data mining in a tax collection agency [6]. Gupta and Nagadevara studied the application of data mining to audit selection strategy and provided a review of some published research articles [19]. Cleary used data mining to help select better targeting cases for audits and further to reduce costs of audits [12]. What distinguishes the work we have done for the pilot project from the work done by others is that we focused on Sales and Use taxes in the USA, and that we not only used data from actual field audits for model training but also reported results from the deployment of the data mining based approach in a real audit project.

Figure 1 illustrates the big picture for the problem addressed in the pilot project: The y-axis is the ratio of the average benefit (or revenue) obtained from an audit case to the average cost of an audit; the x-axis is the number of audits that are ranked according to some criteria, such as the business size. Large businesses are usually selected because they are few in number and typically have higher B/C (Benefit/Cost) values, therefore resulting in higher potential auditing revenue. In contrast, small businesses are selected more or less at random. Improvement of audit selection can be obtained by 1) getting higher B/C values (i.e. creating a lift) in the tail area of this graph, or 2) extending the curve with the same B/C values (i.e. making an extension) in the tail area. In the pilot project, we adopted the first method. That is, we aimed at increasing B/C values for (relatively) small businesses.

In the pilot project, data mining was used to analyze a collection of candidate cases filtered from the database and to identify a smaller set of more profitable candidate cases. The identified candidates would be good cases for field audit. In the pilot project, this was posed as a classification problem, and we used supervised learning techniques with historical data for training.

Fig. 1 The big picture for the audit selection problem

Focusing on a particular tax type and a specific group of taxpayers, in this article we describe the approach used in the pilot project, report validation results, and analyze tax audit data and results collected during a certain period of time. The tax audit data reflects taxpayer behavior and poses certain challenges for data mining. This data contains latent subgroups and suffers from the imperfect nature of real-world data, such as missing values and noise. Models trained on this data are tested using data from actual field audits conducted during a subsequent period of time. It must be noted that the approach used has been validated by the DOR, based on actual field audits performed by auditors (which means that for interested cases, auditors reviewed their business and tax records, visited their business locations, and determined their compliance with the tax laws). The data mining based approach has been used in a real audit project. Thus, this paper presents a unique and valuable case study of a pilot project for mining tax data.

The rest of this paper is organized as follows: Section 2 briefly introduces domain knowledge and background information. Section 3 describes our approach to use data mining in audit selection, and Sect. 4 reports the results for data mining on real-world data. Section 5 reports validation results from actual field audits and Sect. 6 concludes this paper with a discussion on the impact of the pilot project.

2 Background

The DOR is responsible for executing and enforcing the tax laws defined by the legislative process. Enforcement of the tax laws is one key piece of this process. Carrying out audits to identify taxpayers that are furthest from tax compliance is a

central component of this process. The DOR has a limited number of resources to allocate to this process, thus it is of interest to find better ways to identify taxpayers furthest from compliance and therefore allocate resources more efficiently. There are opportunities within the DOR to improve the efficiency of compliance activities and subsequently increase revenue. Data mining can identify these, as demonstrated in the pilot project presented in this article. In looking at the existing compliance efforts, there are essentially two avenues for improving efficiency and increasing revenue: Cost savings and revenue collection.

Cost savings could be understood through the following example. One of the most successful audit projects in the end did not generate sufficient revenue for the DOR. That is, the work on the project returned less than what the DOR invested into the project. If successful in reducing costs for the project, data mining has an even greater potential to reduce costs for audit projects that generate a higher number of unsuccessful audits. Moreover, revenue collection is another essential way to improve efficiency and increase revenue. For every given audit executed, there is potentially a better audit candidate that is not being audited that could provide a higher return for the DOR than the audit this is being performed.

The major issue for audit selection is that of effectively selecting audit cases from a pool of candidates, such that the selected cases will result in substantial revenue gains. Audit selection is the very first step in any tax audit project (no matter whether data mining is used or not). Improving the efficiency of audit selection is a key strategic priority to drive government revenue growth. The more frequently the tax collection agency selects potentially profitable cases for audits, fewer unsuccessful audits could be expected, more cost savings will be achieved, and more revenue will be brought into the government. Audit selection is thus important for all audit projects and requires intensive efforts as well as knowledge from experts.

It is impractical to audit all taxpayers with the limited time and resources provided. Moreover, there is always a cost associated with an audit and the generated revenue might not cover it. Due to these factors, in any audit project experts evaluate certain audit cases and determine taxpayers who are at potential risk for underreporting or underpaying taxes. The final results (i.e. the revenue generated by audit cases) are highly dependent on the quality of the pool of selected taxpayers or audit cases. Although there is a systematic selection approach, it serves more as a guideline and audit cases are generally evaluated by experts based on their experience. Nevertheless, there is room for improvement and data mining has potential to improve the audit selection process.

At the time of the pilot project, the process for audit selection was human-intensive and depended heavily on the experience of experts.³ To begin, rules derived from tax research were used to filter out several thousand candidate cases from a database

³ Some workflows have been changed since the pilot project was completed, and part of the data mining based approach (presented in this article) has been changed because of the introduction of the comprehensive Integrated Tax system, the advances in the analytic capability, and the amount of data available. Nevertheless, the objective of this paper is to share our experiences of using data mining to improve audit selection for the DOR.

or data warehouse (the first stage). This list of candidate cases was then refined and several hundred were selected for field audits (the second stage). In the refinement stage, experts evaluated candidate cases based on pre-specified rules, but evaluation was mostly based on their experience and expertise. If audit cases were well selected in the previous two stages, there was a higher likelihood of generating cost savings and additional revenue. If an audit case turned out to be unsuccessful (i.e. to generate revenue less than the efforts associated with the audit process), the cost was not only time but also the loss of an opportunity to work on a successful case. The goal of the pilot project was to use data mining to improve audit selection, particularly the second stage. As for the first stage, where an initial pool of candidates was generated, the audit selection criteria and process are confidential.

Data mining is a solution that enables increased efficiency and revenue collection and is applicable to most or all of compliance activities. If we view the data mining based approach as a functional component in the whole audit selection process, its input is the initial pool of candidate cases prepared by experts and its output is a collection of labels, each of which corresponds to a case and is with one of the two possible values *Good* or *Bad*. Those labeled as *Good* would be cases for field audit. Nevertheless, data mining is not the only solution to improve audit selection at the DOR.

In the rest of this section, we first introduce some types of taxes, especially Sales and Use taxes. Then, for each tax type we briefly describe the special part in its audit process (used at the time of the pilot project). Given that each tax type has unique characteristics and varying degrees of reliance on data sources other than its own tax return, the audit process varies across divisions at the DOR. All divisions at the DOR are motivated to increase efficiency throughout the audit process, so the approach developed in the pilot project would help the DOR develop the data mining audit processes for other types of taxes.

Sales and Use Taxes Nearly all returns of Sales and Use taxes were processed electronically which allowed the DOR to make use of audit selectors as the return was being processed. Experts used a few dozen selectors. There was room to increase this number but not from the return itself. Once the returns were in the tax system, experts requested worksheets listing subsets of taxpayer financial information. They then sorted this list and added other sources as necessary to look up business taxpayers who were furthest from tax compliance. Audit selection process used both pre-defined queries and ad-hoc queries. However, most audit selection work was delegated to each region to extract independently using the data available in spreadsheets and the related system.

Individual Income Tax Part of the Individual Income Tax audit workload was reduced by filtering out non-compliance that could be identified by looking at the returns alone or in comparison with available federal return data as the return was being processed.

Corporate Franchise Tax Processing the Corporate Tax Returns and getting the paper filed returns into electronic entry format was a difficult task. The returns came

in many different formats and it was a challenge for data entry personnel to fit the data into the standard tax return data structure. The electronic data was therefore somewhat suspect at that time.

Partnership, Estate, Fiduciary, S-Corporation (PEFS) Tax Their filings were entered into the corresponding tax system. Some audit selectors called *audit flags* were implemented. Proven queries and investigative queries played a part in audit selection. Federal, Individual, and other data were merged with PEFS data to identify audit candidates. Additional staff and additional data sources could lead to more investigative audits and ultimately better audit selection quality.

In this paper, we present the work of a pilot study that we have done at the DOR, and the focus was on Sales and Use taxes. The DOR defined the problem addressed in the pilot project a typical binary classification problem to keep the pilot study simple. The reason why the DOR used average cost (and revenue) rather than individual costs (and revenues) was also to keep the pilot study simple. The results, as reported later in this article, showed that the classification models built under such simple settings were beneficial in a real audit project.

3 Approach

Data mining has been applied to finance and accounting [23, 39]. Zhang and Zhou pointed out the challenges of applying data mining to finance [39], and those discussed in this section include choosing data mining techniques, integrating multiple data mining techniques, and using heterogeneous and distributed data sources. The approach developed in the pilot project is based on supervised learning (i.e. classification), while Yang et al. proposed an approach based on unsupervised learning (i.e. clustering) to analyze tax returns [37]. Since we have data sets with labels given by experts at the DOR (and labels themselves are informative), it is reasonable to use supervised learning.

Now let us consider the audit process (for Sales and Use taxes) used at the time of the pilot project. In the final pool of field audits, about half of the audits were fixed irrespective of their outcome. These included the largest companies in every state zone⁴ that were audited regularly. Also included was a group of audits dedicated to research conducted internally by experts. The other half was handpicked by experts using the audit selection process. These hand-picked audits fell under a general category called APGEN (i.e. *Audit Plan—General*), and typically consisted of cases involving small to medium scale sized businesses. Figure 2 presents the process for audit selection for the APGEN category used at the time of the pilot project. Initially, experts posed a database query in order to select several thousand candidate cases out of all the businesses in the state. The query depended on tax type and other

⁴ State zones are divisions of the state, created and used internally by the DOR for efficient work-load balancing.

Fig. 2 The manual audit selection process for audit selection used at the time of the pilot project

information. Additionally, experts could potentially refine the query for candidates satisfying certain other criteria (most likely criteria depending on the state of the economy, industries and other situational factors at the time candidates filled tax returns). Next, experts examined some of the candidates in more detail and selected roughly 10 % for further examination and audit. Experts at the DOR have a method to rank cases, but its confidentiality is protected by law. They would probably also select certain cases based upon suspicions, recommendations from other experts, tip-offs, etc. The final selection was based on case by case subjective evaluations by experts at the DOR. Finally, field audits were conducted on those selected cases and success was measured using audit accuracy along with return on investment (ROI, which represents efficiency and whose definition is given later in this article). In other words, the first two steps in Fig. 2 respectively correspond to the two stages of the process for audit selection (at the time of the pilot project), as introduced in Sect. 1. The more *potentially good* audit cases that were selected in the first two stages, the higher the revenue from the audits. If an audit case turned out to be unsuccessful, then it would be a loss in two respects: (1) time and effort put in the audit was wasted, and (2) the resources could have been directed at a successful case, and thus, potential revenue from a successful audit was lost.

In the pilot project, data mining was applied to the audit selection process (used at the time of the pilot project) and was used to select field audits from the pool of several thousand candidate cases resulting from the database query. That is, the focus of the pilot project was on Step 2 in Fig. 2. As previously mentioned, the selection criteria to generate the initial pool is strictly confidential, thus the goal of the pilot project was to analyze the generated pool rather than to generate such a pool of candidates. Nevertheless, the approach developed in the pilot project could help us ‘data mining practitioners’ and the DOR construct models used to generate the initial candidates. In the pilot project, we focused on Sales and Use taxes and our goal was

to identify candidate cases having higher chances of success of an audit for Sales and Use taxes. The pilot project focused only on audits in the APGEN category (since audits in all other categories were pre-determined), so a binary definition of goodness of an audit was used and defined as the following: Greater than \$500 per year during the audit period is *Good*; less than \$500 per year during the audit period is *Bad*. Audit period is the number of years in the past (starting from the interested fiscal year) for which tax compliance is checked. In almost all cases the audit period is 3 years. In the pilot project, a Use tax assessment of \$1500 was considered a successful audit. This criterion was determined by experts at the DOR. The threshold implied that an audit generating revenue of \$1600 would be viewed as successful as an audit generating revenue of \$16,000. It was what the DOR used to evaluate the selected audits, however. We used it because we intended to make the data mining based approach compatible with the whole audit selection process used by the DOR (at the time of the pilot project). Moreover, it would be useful to consider the cost of individual audits, since the effort to audit a large corporation is more than that to audit a small corporation. Using an average cost for all audits was also what the DOR used in evaluation (at the time of the pilot project). We used it for the same reason: compatibility.

Taxpayer behavior varies across many diverse factors and as a result, even though the focus of the pilot project was on Use tax (while audits for Sales and Use taxes are usually conducted together); data sources from other tax returns were considered as features. Figure 3 illustrates these data sources, including business registration, income, returns of Sales and Use taxes. Use tax field audits and their results conducted over the last 3 years were used to construct the training and test data. Multiple data sources were used for audit selection. This is because business tax data are complicated and related, with certain data sources having potential information regarding Use tax compliance. The training data set consisted of APGEN Use tax audits and their results for the years 2004–2006. The test data consisted of APGEN Use tax audits conducted in 2007 was used to test or evaluate models built on the training data set, while validation was done by actually conducting field audits on predictions made by models built on 2007 Use tax return data (processed in 2008). These three data sets are listed below:

- Training: 2004–2006
- Test or evaluation: 2007
- Validation: 2007

Please note that the two 2007 data sets are different (and there are no common records between the two 2007 data sets).

For data preparation, we started with cleaning the training data set by removing inadequate cases (i.e. cases with none or, at most, one year of tax return data). These cases were generally new businesses or businesses that did not file tax returns, and they had no values (i.e. nulls) for most of the features necessary and were removed from the training data. The data set originally consisting of 11,083 cases was cut down to 10,943 cases after this step. Experts helped us select an initial list of more than 220 features from the various data sources, as shown in Fig. 3. After iterative cycles of

Fig. 3 Data sources for data mining

feature selection and expert consultations a handful of features were selected. These features consisted of two categories: (1) Features related to business characteristics obtained from business registration data, geographic location and type of business, and (2) features correlated with the size of the business, for the three years of the audit period. Nevertheless, details of features that were actually used in the pilot project could not be reported here due to the confidential nature of the tax audit process. Doing so can increase the potential for re-engineering of the audit process, which is clearly undesirable and unlawful.

Initial classification models built using the refined feature set were leaning heavily towards rule-sets that predicted successful audits for larger businesses. During the evaluation stage (using n-fold cross-validation on training data as well as results on test data) it was observed that the initial classification models did well for roughly half the population which consisted of relatively larger businesses, but poorly on the other half which consisted of smaller businesses. The pattern correlating business size with audit success was so dominant that almost all other patterns did not have sufficient relative support to be detected. Therefore, it was decided to divide the original modeling task into two parts: (1) Building one classification model for audit prediction on (relatively) larger businesses (for which the initial classification models seemed to be doing well), and (2) building a second classification model for

(relatively) smaller businesses. We chose to label these two categories as APGEN Large and APGEN Small respectively.

As shown in Fig. 3, we used the same procedure to divide the original three data sets: The training data set was split into an APGEN Large data set and an APGEN Small data set. These two data sets were disjoint. Likewise, the test or evaluation data set was divided into two data sets: one was for APGEN Large and the other was for APGEN Small. The validation data set was also divided into such two data sets.

Businesses in the training set were ranked from largest to smallest, with average annual withholding amount being used as an indicator for business size. The annual withholding amount was directly related to the number of employees and was a very strong indicator of business size. Statistical t-tests were used to determine a withholding amount threshold such that for all businesses below it, there was no significant difference between annual withholding amounts of good and bad audit cases. For cases larger than the threshold value, the business size played an important role in picking good audits (these were mainly the larger businesses in the data set).

The actual withholding amount determined for such a division was determined by using statistical t-tests with various values (to check whether the two sets after a division were really different) and it served as a constant threshold in the pilot project. Thus, the threshold was used to divide the data set into the APGEN Large and APGEN Small categories. Again, the value of the threshold is confidential.

Next, feature selection was performed for APGEN Large and Small individually and different feature sets were obtained for each. Figure 4 presents the feature selection process. Starting from the original feature set, a working feature set was constructed and used as training data to build classification models. We used C4.5 (a decision tree algorithm) [27], Naïve Bayes, multilayer perceptron (an ANN algorithm), support vector machine (SVM) [10, 16], and some others to build classification models. For a set of features in which we were interested, we averaged performance over all the models (excluding those performing significantly poorly). If the results were sufficiently good (on training data and measured using n-fold cross-validation), evaluating the model with test data was performed next. Here, good results were defined as those achieving reasonably high precision and recall with improved estimated ROI (as defined later in this article). In addition, the feature sets corresponding to good results were expected to be consistent with the knowledge and experience of experts. However, at any point where results were not adequately good, the models and their results along with help from experts were examined to identify and remove inadequate features and/or derive new ones. This process was repeated iteratively in order to best use information embedded in the data. Deriving new features was suitable for this purpose and provided a chance for the classification algorithm to analyze the data from different perspectives.

Please note that what is shown in Fig. 4 is not the process used to train models but that used to select features. The test or evaluation data set was used to test or evaluate the feature sets selected to train models, and the results were fed back to the training process. That is, no cases in the test or evaluation data set were used to train a model.

Fig. 4 The feature selection process

We used the same process to select features for APGEN Large and Small, while the set of selected features for APGEN Large was different from that for APGEN Small. Nevertheless, some features were important for both. It is unlawful to reveal the details of the selected features, but we believe that the feature selection process presented in Fig. 4 is valuable and can be used in many other real-world data mining projects.

With the help of experts at the DOR, two special categories of features were recognized for data cleaning and schema reformatting. One special category was composed of features from pre-audit information, which was compiled before auditors performed field audits; the special other category consisted of features from post-audit information, which was collected in or after field audits. However, post-audit information could not be used since it was not available until after auditors performed the field audits. If erroneously used, it could misguide the classification algorithms and created models that succeeded in training (and test or evaluation, probably) but failed in validation or real-world deployment. In fact, any model created using post-audit information was not a predictive model but a descriptive one. While such models were found to be useful to help auditors better understand cases that had already been audited, they were unsuitable for the pilot project's final objective and therefore, not explored further.

The data cleaning process early on filtered out cases that had very little or no data associated with them. The training data set still consisted of quite a few missing values. The data set also consisted of some noise, primarily arising due to errors in reporting from businesses filing their returns or errors due to incorrect data recording from the DOR (especially before the DOR employed the new tax systems). However, this was not a significant issue as the expected fraction of noisy records was very

low. Hence, it was suggested by experts at the DOR to ignore missing values and focus on classification models that were robust towards handling missing values and noisy records.

We reformatted schema (for all data sets) by replacing absolute timestamps with relative timestamps. For example, in the original training data set, three features represented the business statuses of a case in 2004–2006. If a case was audited in 2005, the feature corresponding to its business status in 2006 would have no value. If we intended to use a model in which these three features were referenced to classify a case processed in 2007 (that is, if we intended to *port* the model to a different time frame), we could not find any value in the feature representing the business status of the case in 2004. What is worse, the training data set tracked data only between 2004 and 2006. To solve the problem (or to make the built models *portable*), we used Y, Y-1, and Y-2 for such features. In the first situation given above, there were values for Y and Y-1, and the value for Y-2 was null. In the second situation given above, there were values for Y, Y-1, and Y-2. Then, we could use cases in the training data set even though base values of Y were different. Deligianni and Kotsiantis used a similar way to denote years but for the presentation of the results of forecasting corporate bankruptcy [15].

For modeling, we used WEKA [17, 20], an open source data mining package. Several algorithms had been experimented with and the two with the best performance are reported here. The performance was measured using n-fold cross-validation (on training data and test or evaluation data), and we were looking for (relatively) stable and robust models that would be less prone to overfitting. We intended to have models that would show us the smallest difference between performance on known and performance on unknown data. Having and using such classification models is why and how the data mining based approach performed well in validation where the data set was different from training and test or evaluation data sets. For APGEN Large, MultiBoosting [33] using Naïve Bayes [22, 36] as the base algorithm was used, and for APGEN Small, Naïve Bayes (without MultiBoosting or any other algorithms) was used.

We chose MultiBoosting and Naïve Bayes through experiments, in which we also tried C4.5, multilayer perceptron, SVM, and some others. C4.5 generated interpretable models, but in our experiments it did not demonstrate satisfactory performance. In our experiments, multilayer perceptron and SVM were sensitive to changes in parameters. For both classification algorithms, the best set of parameters obtained on training data was usually different from that obtained on test or evaluation data. Moreover, these two classification algorithms did not perform as well as we expected. In addition, the data sets for APGEN Large and Small might not be sufficiently large for multilayer perceptron and SVM.

Naïve Bayes assumes independence among the features. Despite that this assumption is unrealistic in many situations, classification models built using Naïve Bayes have been successfully used in many real-world applications [28, 36, 38]. Moreover, after we finished the feature selection process, the features we used were less dependent than the original features. MultiBoosting is an ensemble technique that forms a committee (i.e. a group of classification models) to use *group wisdom* to make a

decision. It is different from other ensemble techniques in the sense that it forms a committee of sub-committees (i.e. a group of groups of classification models). Each sub-committee is formed using AdaBoost [18], and wagging (weight aggregation) [2] is used to further combine all these sub-committees into a single committee. It manipulates the given data set to generate different training sets and construct diverse member classification models. The most important characteristic of MultiBoosting is that it exploits the bias reduction capability from AdaBoost as well as the variance reduction capability from wagging. Bias is the average difference between a created model and the underlying model (that generates data and in practice is unknown), while variance represents the average difference among created models. Here the difference arises from using different training sets and it can be measured using error rates. AdaBoost has been shown to have effective bias as well as variance reduction, while it is primarily used for bias reduction. On the other hand, wagging is a variant of bagging [9] (bootstrap aggregation) and is used to reduce variance. Thus, MultiBoosting reduces bias as well as variance [7]. It leverages both techniques and forms a committee that is close to the underlying model and is stable i.e. less variance or more consistent results on unseen data.

Naïve Bayes has high bias and low variance [7, 8]. Thus, it is reasonable to use MultiBoosting with Naïve Bayes (as the base algorithm). Boosting Naïve Bayes was applied to claim fraud diagnosis [31]. Studies showed that MultiBoosting performed well in the prediction of customer choice [30], the prediction of financial distress [26], and the assessment of customer credit quality [21].

4 Evaluation

Classification models were trained or built on tax audit data collected in 2004–2006, while they were tested or evaluated by predicting goodness of audits for 2007 APGEN audit cases. The pilot project used some instead of all such cases because of certain restrictions (mainly those imposed by tax laws). The same withholding amount threshold was used to split the 2007 APGEN data set into (relatively) large and small businesses and the corresponding models were used. The predictions made by models on both large and small businesses were compared to the actual audit results. Figure 5 illustrates the evaluation procedure for data mining based audit selection on the APGEN Large data set for 2007 (which is different from the APGEN Large data set for 2007 used in validation).

APGEN Large for 2007, before the audits⁵, experts predicted 878 cases (out of the initial pool) to be good audits, and after the audits, 495 of them turned out to be good audits. On applying the classification model to the same data set and comparing

⁵ For evaluation, all cases in the APGEN Large and Small data sets were processed and audited. We simulated the audit case selection process and used them as the ground truth. In the simulation, no actual field audits were conducted for the cases that were not selected by experts or the classification models (while we had all the results).

Fig. 5 Evaluation for APGEN Large for 2007 (test or evaluation)

the results to those from actual field audits, it was observed that the classification model predicted 534 cases (out of the initial pool) to be good audits, out of which 386 cases (or 72.3 %) were actually good audits. To sum up, for the same pool of candidate cases for APGEN Large for 2007, experts suggested to audit 878 cases (and 495 of them were good), while the classification model suggested to audit 534 cases (and 386 of them were good). Results were evaluated on the ROI metric. ROI is used as a measure of efficiency, as shown in Eq. 1, which is suggested by experts at the DOR.

$$\text{Efficiency} = \text{ROI} = \frac{\text{Total revenue generated}}{\text{Total collection cost}}. \tag{1}$$

Simplicity and compatibility are the reasons why we did not use ROC (receiver operating characteristic) as the performance measure. ROC was not used by experts at the DOR. Moreover, in some audit projects, individual results were not be available (while only average results were available) and hence ROC was not applicable.

Business analysis is important for any practical data mining application. Cleary and Tax reported their experience in using data mining to help select better targeting cases for audits with little business analysis [12]. In the pilot project, we followed the suggestions given by experts at the DOR to perform business analysis and reported the results as follows.

Table 1 summarizes results from the manual audit selection process at the time of the pilot project, while Table 2 presents results using the data mining based approach. In both tables, the first row (disregarding the header) indicates the number of audits (and corresponding dollar amounts) that were selected by the process. For evaluation purposes the following were used (these were estimates suggested by experts and not the actual numbers): the average number of hours spent conducting a Use tax audit is 23 (h), while the average pay of a tax specialist is \$20 per hour. Thus, the collection cost of k audits was \$460k. The value seems low, but this was how experts

Table 1 Business analysis for the manual audit selection process used at the time of the pilot project for APGEN Large for 2007 (test or evaluation)

	Good audits	Bad audits	Total
Number of audits	495	383	878
	(56.4 %)	(43.6 %)	(100 %)
Revenue generated	\$6,502,724	\$170,849	\$6,673,573
	(97.4 %)	(2.6 %)	(100 %)
Collection cost	\$227,700	\$176,180	\$403,880
	(56.4 %)	(43.6 %)	(100 %)

Table 2 Business analysis for the classification model created for APGEN Large for 2007 (test or evaluation)

	Good audits	Bad audits	Total
Number of audits	386	148	534
	(72.3 %)	(27.7 %)	(100 %)
Revenue generated	\$5,577,431	\$72,744	\$5,650,175
	(98.7 %)	(1.3 %)	(100 %)
Collection cost	\$177,560	\$68,080	\$245,640
	(72.3 %)	(27.7 %)	(100 %)

at the DOR suggested to calculate the cost for the pilot project. In Tables 1 and 2, the second and the third rows report revenue generated and collection cost, respectively.

The ROI value for the manual audit selection process used at the time of the pilot project for APGEN Large is 1652 % and the same for the data mining based approach is 2300 %. This represents a 39.2 % increase in efficiency. Figure 6 illustrates audit resource deployment efficiency (as does Fig. 8). In Fig. 6, the x-axis and y-axis respectively represent the number of audits performed (i.e. audit effort) and the number of audits that are successful and generate revenue. Theoretically, the best situation is that all audits performed turn out to be profitable. This is captured by the left-most (solid) line. Additionally, the manual audit selection process is represented by the right-most (solid) line. Based on these two lines, the space is divided into three regions, as shown in both figures: the left-most region (A) represents the impossible situation of having more successful audits than the actual number of audits conducted. The right-most region (C) represents the situation where, compared to the manual audit selection process, fewer successful audits are obtained. Any model in this region is less efficient than the manual audit selection process. The data mining based approach is located in the region B. Any solution in region B represents a method better than the audit selection process and is closer to the theoretically best process.

From Fig. 6, the theoretically best process will find 495 successful audits when 495 audits are performed, while the manual audit selection process used at the time of the pilot project will need 878 audits in order to obtain the same number of successful audits. If one projects the (solid) line presenting the data mining based approach, it is observed that in order to obtain 495 successful audits, the number of audits performed will be lower than 878 (which is better than the manual audit

Fig. 6 Audit resource deployment efficiency for APGEN Large for 2007 (for Tables 1 and 2)

Table 3 The confusion matrix for APGEN Large for 2007 (test or evaluation); R revenue, C collection cost

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = \$5,577,431 (83.6 %) C = \$177,560 (44 %)	109 (Use tax lost) R = \$925,293 (13.9 %) C = \$50,140 (12.4 %)
Actually bad	148 (costs wasted) R = \$72,744 (1.1 %) C = \$68,080 (16.9 %)	235 (costs saved) R = \$98,105 (1.4 %) C = \$108,100 (26.7 %)

selection process). It can be estimated as $534 \times 495 \div 386$, which is approximately 685. Alternatively, if the manual audit selection process selects only 534 cases, the number of successful audits will be lower than 386, which is found by the data mining based approach. It can be estimated as $495 \times 534 \div 878$, which is approximately 301. The former number shows that with data mining, less effort is required for the same degree of tax compliance, while the latter number shows that higher tax compliance is achievable for the same effort. Furthermore, Table 3 presents the confusion matrix for the classification model on the APGEN Large data set (for 2007). Columns and rows are for predictions and actual results, respectively. Revenue and collection cost associated with each element is also reported. The top-left element indicates Use tax assessment collected, the top-right element indicates Use tax assessment lost (i.e. cases predicted as bad turning out to be good), the bottom-left element indicates collection costs wasted due to audits incorrectly predicted as good, and the bottom-right element indicates collection costs saved when predicted bad audits are not assessed. Notice that the data mining based approach eliminated cases that

Fig. 7 Evaluation for APGEN Small for 2007 (test or evaluation)

Table 4 Business analysis for the manual audit selection process used at the time of the pilot project for APGEN Small for 2007 (test or evaluation)

	Good audits	Bad audits	Total
Number of audits	99 (20.9 %)	374 (79.1 %)	473 (100 %)
Revenue generated	\$527,807 (98.7 %)	\$93,259 (1.3 %)	\$621,066 (100 %)
Collection cost	\$45,540 (20.9 %)	\$172,040 (79.1 %)	\$217,580 (100 %)

consumed 26.7 % of collection resources but generated only 1.4 % of revenue, thus, significantly improving efficiency.

Figure 7 illustrates the audit selection process used at the time of the pilot project and the data mining based approach for APGEN Small. Here, experts predicted 473 cases as good audits. After conducting field audits, only 99 of them were actually good. For businesses in this subgroup, only one-fifth of cases selected by the audit selection process generated revenue greater than the pre-defined threshold value. In contrast, 47 out of 140 cases (33.6 %) selected by the classification model were truly good audits.

Table 4 reports results for the manual audit selection process used at the time of the pilot project, and Table 5 summarizes results for the data mining based approach. Apart from the increase in precision, the ROI value for the manual audit selection process for APGEN Small is 285 % and the same for the data mining based approach is 447 %, indicating a 57 % increase in efficiency.

Similar to Fig. 6, Fig. 8 illustrates audit resource deployment efficiency. From Fig. 8, when 99 audits are performed the theoretically best process will find 99 profitable audits. However, the manual audit selection process will need 473 audits in order to obtain the same number of successful audits. For using the data mining based approach to obtain 99 successful audits, the number of audits performed will

Fig. 8 Audit resource deployment efficiency for APGEN Small for 2007 (for Tables 4 and 5)

Table 5 Business analysis for the classification model created for APGEN Small for 2007 (test or evaluation)

	Good audits	Bad audits	Total
Number of audits	47 (33.6 %)	93 (66.4 %)	140 (100 %)
Revenue generated	\$263,706 (91.5 %)	\$24,441 (8.5 %)	\$288,147 (100 %)
Collection cost	\$21,620 (33.6 %)	\$42,780 (66.4 %)	\$64,400 (100 %)

be lower than 473. It can be estimated as $140 \times 99 \div 47$, which is approximately 295. This number shows that with data mining less effort is required to obtain the same degree of tax compliance. Furthermore, 47 out of 140 cases selected by the data mining based approach turn out to be successful audits. If the manual audit selection process is used to select 140 audits, the number of successful audits will be lower than 47. It can be estimated as $99 \times 140 \div 473$, which is approximately 29. This number shows that with data mining higher tax compliance is achievable for the same effort.

The confusion matrix for the classification model for APGEN Small is presented in Table 6. The 47 good audits correctly identified correspond to cases that consume 9.9 % of collection costs but generate 42.5 % of revenue. Note that the 281 bad audits correctly predicted by the classification model represent notable collection cost savings. These are associated with 59.4 % of collection costs generating only 11.1 % of the revenue.

Table 6 The confusion matrix for APGEN Small for 2007 (test or evaluation); *R* revenue, *C* collection cost

	Predicted as good	Predicted as bad
Actually good	47 (Use tax collected) R = \$263,706 (42.5 %) C = \$21,620 (9.9 %)	52 (Use tax lost) R = \$264,101 (42.5 %) C = \$23,920 (11 %)
Actually bad	93 (costs wasted) R = \$24,441 (3.9 %) C = \$42,780 (19.7 %)	281 (costs saved) R = \$68,818 (11.1 %) C = \$129,260 (59.4 %)

Fig. 9 Validation for the data mining based approach

5 Validation

This section reports results from actual field audits conducted by auditors at the DOR. In order to evaluate the pilot project, the DOR validated the data mining based approach by using them to select cases for actual field audits in a real audit project. The audit selection criteria and process to generate the initial pool of candidate cases are confidential due to DOR regulations. Thus, the data mining based approach was used to select cases from the pool for actual field audits. Figure 9 illustrates the process used to validate the data mining based approach.

The DOR used the classification models to select 414 tax cases for which auditors conducted actual field audits. For the pilot project, the DOR defined a productive audit as an audit resulting in an assessment of at least \$500 per year, or \$1500 per case (for a 3-year audit period). The classification models were used to analyze the collected tax data and the top 414 most likely predicted good audits were selected. For these selected cases, auditors reviewed their business and tax records, visited

Table 7 Validation results in success rate

	Pre data mining (%)	Data mining predicted (%)	Actual (%)
Sales	29 %	38 %	37 %
Use	39 %	56 %	51 %

Table 8 Validation results

	Pre data mining (\$)	Data mining predicted (\$)	Actual (\$)
Sales	6497	11,976	8186
Use	5019	8623	10,829

their business locations, and determined their compliance with the tax laws. The data mining approach used in the pilot project had been used in a real audit project and therefore the DOR did not carry out the manual audit selection process in parallel. Consequently, there was no direct comparison between the data mining based approach used here and the manual audit selection process. However, as shown in the last step of Fig. 9, audit results from the data mining based approach were reviewed by experts.

The DOR reviewed the results of the actual field audits and compared them to the predicted ones. Table 7 reports results in success rate (i.e. accuracy) while Table 8 reports results in dollars. Both tables present results for Use tax and Sales tax even though the focus of the pilot project was on Use tax (as in earlier discussion). On the one hand, auditors would simultaneously do Use tax and Sales tax when they decided to audit a taxpayer, no matter if the decision is based on their analysis for Use tax or Sales tax for the taxpayer (and no matter whether data mining is used or not). On the other hand, auditors would probably concentrate on Sales tax even though the initial decision was from their analysis of Use tax. Such a decision depended on their experience, the possible collection cost, and the potential profits of these selected audits.

As one can see from Table 7, only 29 % of audits for Sales tax were thought to be profitable by the manual audit selection process (used at the time of the pilot project), named *pre data mining* process, while 38 % of audits were predicted as profitable by the data mining based approach. After auditors performed actual field audits, 37 % of audits turned out to be successful and generated revenue for the DOR. Similarly, 56 % of audits for Use tax were predicted as profitable by the data mining based approach, while only 39 % of audits were thought to be profitable by the manual audit selection process. For Use tax, the actual success rate was 51 %, which was closer to the rate predicted by the data mining based approach. These numbers validated that the data mining based approach had better accuracy and consequently better efficiency.

Table 8 reports revenue in dollars predicted by the manual audit selection process (used at the time of the pilot project) and by the data mining based approach, and it also reports the revenue actually collected by auditors after they performed actual field audits. The *pre data mining average* dollars collected was from historical data (e.g. from auditors' experience) but not from conducting the manual audit selection

Table 9 Validation results of 414 actual filed audits for different categories

	Overall total assessed (\$)	Overall average assessed (\$)
Large use and sales	1,399,436	19,437
Small use and sales	72,605	2504
Large sales	6,229,248	23,776
Small sales	101,895	1998
Combined totals	7,803,184	18,848

process in parallel. Experts provided qualitative assessment of the cases that were selected by the data mining based approach before auditors performed actual field audits, as the column *data mining predicted* suggests. The results from actual field audits were in the last column. The detailed assessment results are not reported here since they are protected by law. Results in dollars for different categories are reported in Table 9. As described earlier, auditors would concentrate their attention on Use tax, Sales tax, or both once they decided to conduct field audits. Therefore, there are different categories shown in Table 9. These results clearly show that the data mining based approach generated more revenue for the DOR. For example, in Table 9, the DOR assessed Sales tax of \$23,776 in average for relatively large businesses. Please recall that the threshold for being a profitable audit was set to \$1500 per case (for a 3-year audit period). The result achieved by the data mining based approach clearly proved that it was able to not only save costs and efforts but also generate more revenue. What is more, the manual audit selection process (used at the time of the pilot project) was struggling with relatively small businesses and usually most cases generated less than \$1500. Nevertheless, the average amount of assessed Sales and Use taxes achieved by the data mining based approach was \$2504. Furthermore, if auditors decided to concentrate on Sales tax, the average assessed amount for relatively large businesses was \$23,776 while that for relatively small businesses was \$1998. Considering the threshold was set to \$1500 per case, and the average assessed amount of dollars was \$18,848, revenue generated by the data mining based approach was over 12 times of the threshold that was associated with the average collection cost. These results demonstrated that data mining had the potential to efficiently and effectively perform more sophisticated tax audit selection.

6 Conclusions

We have presented a case study of a pilot project and shared our experiences of using data mining to improve audit selection for the Minnesota Department of Revenue (DOR). Additionally, we have described some practical challenges when applying data mining to audit selection. Improving the efficiency of audit selection and further the productivity of the tax assessment process is an essential component of driving revenue growth for the DOR as well as the government. The audit selection process

is a human-intensive process required of knowledgeable experts. However, apart from being cumbersome, the process is also inefficient. Bad audits not only waste auditors' time and resources but also erode revenue. The approach developed in the pilot project is to use the data mining based approach (i.e. classification models) for the purpose of improving audit selection. Since data play a vital role in any data mining project, considerable attention was paid to data pre-processing, cleaning, and reformatting. Models were trained and tested using real-world data. The results of the pilot project showed that the data mining based approach achieved an increase of 63.1 % in efficiency. The most important part of the pilot project is the validation from actual field audits which demonstrated the usefulness of data mining for improving audit selection in terms of accuracy as well as revenue generated. Improving government efficiency is important for effective governance, while improving tax assessment efficiency is essential for economic activities. This is especially critical in a tough economy. The pilot project provided a further impact of increased interest among the government for effective applications of data mining. The direct impact is a reexamination and refinement of other tax assessment processes that are in use but may be inefficient.

Acknowledgements The research was supported in part by a grant from the Minnesota Department of Revenue. This collaboration would not have been possible without the sponsorship and support of a number of organizations and individuals. Specifically, the authors would like to thank the Minnesota Department of Revenue and the University of Minnesota for providing the institutional support for this work. Finally, we would like to acknowledge the support of the Office of Enterprise Technology, State of Minnesota, and the vision and leadership of Commissioner Gopal Khanna for creating a novel relationship which allows University faculty and students to interact with government departments to mutually collaborate on using advanced technologies to address the State of Minnesota's information technology needs. The opinions expressed in this article, however, are solely those of the authors, and do not represent, directly or by implication, the policies of their respective organizations. The authors would also like to thank anonymous reviewers for their valuable comments and suggestions.

References

1. Andreoni, J., Erard, B., Feinstein, J.: Tax compliance. *J. Econ. Lit.* **36**(2), 818–860 (1998)
2. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learn.* **36**(1), 105–139 (1999)
3. Bhowmik, R.: Detecting auto insurance fraud by data mining techniques. *J. Emerg. Trends Comput. Inf. Sci.* **2**(4), 156–162 (2011)
4. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D.: A classification-based methodology for planning audit strategies in fraud detection. In: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, pp. 175–184 (1999)
5. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D.: Using data mining techniques in fiscal fraud detection. In: *Proceedings of the 1st International Conference on Data Warehousing and Knowledge Discovery*, Florence, Italy, pp. 369–376 (1999)
6. Bots, P.W.G., Lohman, F.A.B.: Estimating the added value of data mining: A study for the Dutch Internal Revenue Service. *Int. J. Technol. Policy Manag.* **3**(3/4), 380–395 (2003)

7. Brain, D., Webb, G.I.: On the effect of data set size on bias and variance in classification learning. In: *Proceedings of the 4th Australian Knowledge Acquisition Workshop*, Sydney, Australia, pp. 117–128 (1999)
8. Brain, D., Webb, G.I.: The need for low bias algorithms in classification learning from large data sets. *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, Helsinki, Finland, pp. 62–73 (2002)
9. Breiman, L.: Bagging predictors. *Machine Learn.* **24**(2), 123–140 (1996)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3), 27 (2011)
11. Chen, Y.S., Cheng, C.H.: A Delphi-based rough sets fusion model for extracting payment rules of vehicle license tax in the government sector. *Expert Syst. Appl.* **37**(3), 2161–2174 (2010)
12. Cleary, D.: Predictive analytics in the public sector: Using data mining to assist better target selection for audit. *Electron. J. e-Gov.* **9**(2), 132–140 (2011)
13. Cornia, G.C., Sjoquist, D.L., Walters, L.C.: Sales and use tax simplification and voluntary compliance. *Public Budget. Financ.* **24**(1), 1–31 (2004)
14. DeBarr, D., Eyler-Walker, Z.: Closing the gap: Automated screening of tax returns to identify egregious tax shelters. *ACM SIGKDD Explor. Newslett.* **8**(1), 11–16 (2006)
15. Deligianni, D., Kotsiantis, S.B.: Forecasting corporate bankruptcy with an ensemble of classifiers. In: *Proceedings of the 7th Hellenic Conference on Artificial Intelligence*, pp. 65–72 (2012)
16. EL-Manzalawy, Y., Honavar, V.: WLSVM: Integrating LibSVM into Weka environment. <http://www.cs.iastate.edu/~yasser/wlsvm> (2005). Accessed 17 Feb 2012
17. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H., Trigg, L.: Weka—A machine learning workbench for data mining. In: Maimon, O., Rokach, L. (eds.) *Data Mining and Knowledge Discovery Handbook*, pp. 1269–1277. Springer, Berlin (2010)
18. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: *Proceedings of the 13th International Conference on Machine Learning*, pp. 148–156 (1996)
19. Gupta, M., Nagadevara, V.: Audit selection strategy for improving tax compliance—Application of data mining techniques. In: Agarwal, A., Venkata Ramana, V. (eds.) *Foundations of E-government*. Computer Society of India, Hyderabad (2007)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* **8**(1), 10–18 (2009)
21. Huang, S.C., Wu, C.F.: Customer credit quality assessments using data mining methods for banking industries. *Afr. J. Bus. Manag.* **5**(11), 4438–4445 (2011)
22. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 338–345 (1995)
23. Kirkos, E., Manolopoulos, Y.: Data mining in finance and accounting: A review of current research trends. In: *Proceedings of the 1st International Conference on Enterprise Systems and Accounting*, pp. 63–78 (2004)
24. Kirkosa, E., Spathisb, C., Manolopoulosc, Y.: Data mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.* **32**(4), 995–1003 (2007)
25. Kotsiantis, S., Koumanakos, E., Tzelepis, D., Tampakas, V.: Forecasting fraudulent financial statements using data mining. *Int. J. Comput. Intell.* **3**(2), 104–110 (2006)
26. Liu, H., Huang, S.: Integrating GA with boosting methods for financial distress predictions. *J. Qual.* **17**(2), 131–158 (2010)
27. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco (1993)
28. Rish, I.: An empirical study of the naïve bayes classifier. Tech. rep., IBM. <http://researchweb.watson.ibm.com/people/r/rish/papers/RC22230.pdf> (2001). Accessed 17 Feb 2012
29. Toder, E.: Reducing the tax gap: The illusion of pain-free deficit reduction. Tech. rep., Tax Policy Center. http://www.taxpolicycenter.org/UploadedPDF/411496_reducing_tax_gap_revised.pdf (2007). Accessed 17 Feb 2012
30. van Wezel, M., Potharst, R.: Improved customer choice predictions using ensemble methods. *Eur. J. Oper. Res.* **181**(1), 436–452 (2007)

31. Viaene, S., Derrig, R.A., Dedene, G.: A case study of applying boosting Naïve Bayes to claim fraud diagnosis. *IEEE Trans. Knowl. Data Eng.* **16**(5), 612–620 (2004)
32. Wang, J., Yang, J.G.S.: Data mining techniques for auditing attest function and fraud detection. *J. Forensic Invest. Account.* **1**(1) (2009). <http://www.bus.lsu.edu/accounting/faculty/lcrumbley/jfia/Articles/FullText/2009v1n1a8.pdf>
33. Webb, G.I.: Multiboosting: A technique for combining boosting and wagging. *Machine Learn.* **40**(2), 159–196 (2000)
34. Webley, P., Cole, M., Eidjar, O.P.: The prediction of self-reported and hypothetical tax-evasion: Evidence from England, France and Norway. *J. Econ. Psychol.* **22**(2), 141–155 (2001)
35. Wu, R.C.F.: Integrating neurocomputing and auditing expertise. *Manag. Audit. J.* **9**(3), 20–26 (1994)
36. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**(1), 1–37 (2008)
37. Yang, Y., Ge, E., Barns, R.: Towards effective and efficient identification of potential tax agent compliance risk: A stratified random sampling approach. *e-J. Tax Res.* **9**(1), 116–137 (2011)
38. Zhang, H.: The optimality of naïve Bayes. In: *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, Miami Beach, FL, USA (2004)
39. Zhang, D., Zhou, L.: Discovering golden nuggets: Data mining in financial application. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **34**(4), 513–522 (2004)