TEDM-PU :A Tax Evasion Detection Method Based on Positive and Unlabeled Learning

1st Yingchao Wu
SPKLSTN Lab, School of EIE
Xi'an Jiaotong University
Xi'an, China
xjtuwuyc@163.com

2nd Qinghua Zheng
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
bodong@mail.xjtu.edu.cn

3rd Yuda Gao
SPKLSTN Lab, School of EIE
Xi'an Jiaotong University
Xi'an, China
liana_gyd@163.com

4th Bo Dong
Collage of Distance Education
Xi'an Jiaotong University
Xi'an, China
dong.bo@mail.xjtu.edu.cn

5th Rongzhe Wei
SPKLSTN Lab, School of EIE
Xi'an Jiaotong University
Xi'an, China
jessonwrz@163.com

6th Fa Zhang
SPKLSTN Lab, School of EIE
Xi'an Jiaotong University
Xi'an, China
724183028@qq.com

7th Huan He
SPKLSTN Lab
Xi'an Jiaotong University
Xi'an, China
hehuan@mail.xjtu.edu.cn

***Abstract*—Tax evasion detection plays a crucial role in reducing tax revenue loss and many efforts have been made to develop detection models based on machine learning techniques. To train an effective model to detect tax evaders, a large amount of data is required, especially sufficient labeled data. However, the expensive and time-consuming annotation process results in small amount of labeled data being available, which makes the development of detection models difficult. To address this issue, we propose a tax evasion detection method based on positive and unlabeled learning (TEDM-PU), to identify tax evasion by utilizing limited annotated tax evasion taxpayers and a large amount of unlabeled data. The TEDM-PU framework consists of three stages: a preprocessing stage extracting taxpayer features based on random forest, a pseudo labeling stage assigning pseudo labels to unlabeled samples based on PUAdapter, and a model training stage based on LightGBM method. To evaluate the effectiveness of our proposed TEDM-PU, we conduct experimental tests on real-world tax data. The results demonstrate that TEDM-PU method can detect tax evaders with higher accuracy and better interpretability than state-of-the-art methods.**

***Index Terms*—tax evasion detection, PU learning, interpretability**

I. INTRODUCTION

Tax is one of the most important types of fiscal revenue, and with which the government can regulate the allocation of financial resources and the distribution of national incomes [1]. However, in order to retain more profits, many companies adopted illegal means to evade taxes, such as reducing taxable income and transferring assets [2]. Especially in recent years, tax evasion methods have become more diverse and covert. Some companies use advanced facilities, accounting methods, and human factors to evade taxation inspections [10], which makes auditing work more difficult. These tax evasions can cause massive tax loss in national fiscal revenue. For instance, the State Taxation Administration of China reported that the tax revenue loss in China accounted for 9.99% of its gross national product in 2013 [1].

In order to detect tax evasion, a large number of methods have been used and studied, including the following three types: manual case selection, whistle-blowing-based case selection, and computer-based case selection [2]. The manual case selection and whistle-blowing-based case selection methods mainly rely on the personal experience of tax experts to identify suspicious tax evaders from tax data, which cannot meet the demand of tax detection in large-scale tax data. Therefore, the computer-based case selection methods are being widely used in recently studies and tax evasion detection systems [2]. The computer-based case selection extracts evasion-related features from historical data and trains tax evasion detection model based on machine learning methods, which can be considered a semi-automatic and labor-saving method [10].

However, due to the complexity of tax data and the lack of tax experts, the data annotation in the tax domain is very expensive and time consuming. For example, the number of annual tax-related business records has exceeded 1 billion, and the daily peak of these records is up to 10 million in China [10]. Given limited staff resources and such huge amount of tax data, the number of tax evaders that tax experts can identified is very limited. As a result, there is no sufficient labeled tax data available for training an effective detection mode, which limits the use of the computer-based case selection method in real productive environments. Therefore, how to construct a unified tax evasion detection model with a small amount of tax evader data and a large amount of unlabeled data remains a pressing issue.

To address this issue, we apply positive and unlabeled learning (PU Learning) method [5] to utilize the large amount of unlabeled tax data in the tax evasion detection. PU learning is a semisupervised binary classification model that trains a binary classifier through positive samples and unlabeled samples. The goal of PU learning is the same as general binary classification: to learn a model that is able to distinguish between positive and negative samples. PU learning reduced the need for annotated data and solved the problem of data imbalance. It has been

widely applied in disease gene identification [6], knowledge base construction [7], text classification [8] [9], etc.

However, there are three challenges when applying PU learning to tax evasion detection.

First, to the best of our knowledge, there are no studies on tax evasion detection based on PU learning. Few works have explored how PU learning can be used in tax evasion detection. Current studies about PU learning are difficult to directly apply to tax evasion detection due to the high accuracy and interpretability requirements in the field of taxation.

Second, as presented by Tian et al. [2], the results of machine learning-based methods are not explainable and counterintuitive. Almost all machine learning models and PU learning methods are black box models due to the feature mapping operation, which is vulnerable to security attacks [4] [5] [6]. Making the model interpretable is an important issue for developing a robust and stable tax evasion detection system.

Third, due to the complexity of tax data, the number of tagged tax evaders is very small. This small number of tax evaders is also manually labeled by tax experts and is not entirely correct, which leads to difficulties in the PU learning process.

To address the above challenges, we propose A Tax Evasion Detection Method Based on Positive and Unlabeled Learning (TEDM-PU) method for tax evasion detection task in positive and unlabeled tax data, which consists of three stages: 1) a preprocessing stage extracting taxpayer features based on random forest method [37], 2) a pseudo labeling stage assigning pseudo labels to unlabeled samples based on PUAdapter [11], and 3) a model training stage based on LightGBM [12] method. Specifically, a random forest classifier is used to select the most useful information from the feature space. PUAdapter is a PU learning method, that ensures TEDM-PU can identify reliable positive and negative samples from unlabeled samples. We use a feedforward artificial neural network (ANN) [13] that consists of three layers of nodes; each node is a neuron that uses a nonlinear activation function, that classify data that PUAdapter cannot clearly distinguish. LightGBM supports the interpretability and accuracy of the TEDM-PU. Therefore, the TEDM-PU can provide an effective and explainable model for tax evasion detection task in positive and unlabeled tax data.

To evaluate the effectiveness of the TEDM-PU, we conduct experiments on real world tax dataset collected from local taxation administration. The results show that the TEDM-PU can detect tax evaders with higher accuracy and better interpretability than state-of-the-art methods

The main contributions of this paper can be summarized as follows:

- We propose a new tax evasion detection method by considering real tax data that only include a small amount of tax evaders and a large amount of unlabeled data and the interpretability of the derivation results.
- We propose a novel tax evasion detection method, namely TEDM-PU, which integrates PUAdapter, ANN and LightGBM to detect tax evasion in large-scale tax data.
- We justify the performance of the TEDM-PU through comparison with existing work based on a large real-world dataset . The results show that the TEDM-PU can greatly improve the accuracy of tax evasion detection and provide better interpretability than existing work.

The remainder of the paper is organized as follows. Section 2 provides a brief review on related work. In Section 3, we formulate the problem and lists key notations. We propose the TEDM-PU framework and its details in Section 4. We describe the experimental results and provide analysis and discussions in Section 5. Finally, a conclusion is presented in the last section.

II. RELATED WORK

In this section, we briefly review the related work on tax evasion detection and PU learning methods.

A. Tax Evasion Detection Methods

Many novel techniques for detecting financial fraud have been proposed in the literature. The current auditing methods used by tax authorities include: manual case selection, whistle-blowing-based selection and computer-based case selection methods. Manual case selection and report-based selection methods rely on the experience of tax auditors and are time-consuming and labor-intensive. Therefore, computer-based case selection methods are the most effective method used by tax authorities to detect tax evasion.

Existing machine learning-based tax evasion detection methods include association analysis [14], cluster analysis [15] [16] [17] [18], classification [19] [20] [21] [22], simulation [23] [24] [25], reinforcement learning [26] [27] and transfer learning [10]. For example, Liu et al. [17] used hierarchical clustering in tax inspection case selection based on seven financial indexes. Wu et al. [14] employed association rules to a value-added tax database to uncover patterns and relationships among attributes that are useful for identifying tax evasions. Chen and Cheng [19] proposed a hybrid model that combines a Delphi method and a rough set classifier to classify vehicle license tax payment. Abe et al. [26] developed a constrained Markov decision process-based approach to the problem of optimally tax management, general debt, and collections processes at financial institutions. Zhu et al. [10] used transfer learning to construct an evasion detection model for a region with the help of auxiliary data from another region.

However, there are two main problems of machine learning-based tax evasion detection methods. First, they use statistical techniques to identify whether a taxpayer has evaded taxes and require a set of manually labeled data contributed by auditors. Because data annotations in the tax field are very time consuming, there are very few labeled data. Second, it is difficult to explain and trace the results obtained by applying these models.

B. PU Learning Methods

PU learning methods can be divided into the following three categories: two-step techniques [29] [30] [31], biased

learning [32] [33] [34] and class prior incorporation [11] [35] [36]. The two-step technique consists of two steps: 1) identifying reliable negative examples, and 2) learning based on the labeled positives and reliable negatives. Biased learning considers PU data as fully labeled data with class label noise for the negative class. Class prior incorporation modifies standard learning methods by applying the mathematics from the SCAR assumption directly, using the provided class prior. Our method is related to class prior incorporation PU learning. The motivation is that: class prior incorporation PU learning avoids tuning the weights, classifies positive and unlabeled data very well, preserves the original tax features and enhance the interpretability of the tax evasion detection model.

PUAdapter [11] is the most classic method used to solve the PU learning problem through class prior incorporation. It uses the "completely random selection" assumption to construct a high-quality classification model for positive and unlabeled data. R. Kiryo et al. [41] noted that it might lead to unbiased risk estimators if we unrealistically assume that the class-posterior probability is one for all positive data. Hence R. Kiryo et al. proposed a method named nnPU [41] to address this limitation. However, class prior incorporation PU learning works poorly when the positive instance size is small because it needs to calculate the a priori value with positive instances.

Regarding the PU learning problem, the aim is to build an accurate binary classifier without the need to collect negative examples for training. A two-step [30] approach is a solution for the PU learning problem that consists of two steps: (1) identifying a set of reliable negative documents, and (2) building a classifier iteratively. F. Mordelet et al. [39] propose a new method for PU learning with a conceptually simple implementation based on bootstrap aggregating (bagging) techniques. The algorithm iteratively trains many binary classifiers to discriminate the known positive examples from random subsamples of the unlabeled set, and averages their predictions. However these methods need to constantly adjust the weights.

Although these techniques have been applied and examined in different domains, studies on tax evasion detection based on PU learning are lacking. We construct a tax evasion detection model by applying PU learning, using a small list of tax evaders and a large amount of unlabeled tax audit data. Based on our survey, no study has assessed how to use PU learning to solve the problem of insufficient data in tax audits to date.

III. PROBLEM STATEMENT AND NOTATIONS

In this paper, taxpayers who evade taxes are treated as positive samples, and normal taxpayers are treated as negative samples. We focus on three main problems that are widespread in tax evasion detection. (i) how to build an accurate tax evasion detection model for tax data with only a small number of positive training samples and a large number of unlabeled training samples, (ii) how to solve the problem that there are some errors labels in the labeled positive samples, and (iii) how to induce interpretability in a tax evasion detection model?

To address the above challenges, this paper propose a Tax Evasion Detection Method Based on Positive and Unlabeled Learning. We calculate the probability value that each sample belongs to the positive sample using PUAdapter. Then, each sample is given a pseudo label using thresholds h_1,h_2 and ANN. Finally, a tax evasion detection model is built with source data and pseudo labels. We combine PU learning with an interpretable classifier to induce interpretability in the tax evasion detection model.

Next, we provide some notations and definitions.

We denote the tax data set as X. Here $P = \{x_1, ..., x_m\} \subset X$ with a positive label, and $U = \{x_1, ..., x_n\} \subset X$ without any labels. Let $y \in \{-1, 1\}$ be a binary label, where $y = 1$ represents taxpayer tax evasion and $y = -1$ represents a normal taxpayer. Let $s = 1$ if the sample x is labeled, and let $s = 0$ if x is unlabeled. Only positive sample are labeled. Thus, $y = 1$ is certain when $s = 1$, but when $s = 0$. Then either $y = 1$ or $y = -1$ may be true. Our target is to use P and U to learn a tax evasion detection model $f : x \to y$.

We summarize the notations used in the paper in Table I.

TABLE I
NOTATIONS USED

Variable	Description
X	Tax Audit DataSet
Y	Label set of X
S	A set of indicator variables for samples to be labeled
x	A set of vectors of attributes of samples
y	Label of the x
s	Indicator variable for an samples to be labeled
P	Positive data set
U	Unlabeled data set
n	Number of instances in the positive data set
m	Number of instances in the unlabeled data set
q_j	j^{th} term in the feature space

IV. PROPOSED SCHEME

In this section, we describe the framework of the TEDM-PU, the details of our method, and the TEDM-PU procedure.

A. Framework of the TEDM-PU

The TEDM-PU is a novel tax evasion detection method based on PU learning, which aims to construct an evasion detection model for tax data with only a small number of positive and a large number of unlabeled samples. It provides a unified framework and builds interpretable classifiers to handle tax evasion detection problems with only positive and unlabeled samples.

The framework of the TEDM-PU is shown in Figure 1 and is composed of three main stages.

In the preprocessing stage, the tax audit dataset consists of a small number of positive and a large number of unlabeled samples. They are both high-dimensional semi-structured data sets. Therefore, preprocessing is required for tax data, including filling in missing values, reducing feature dimensions and extracting features from sequence data. A random forest method [37] is adopted to automatically select the features associated with tax evasion.

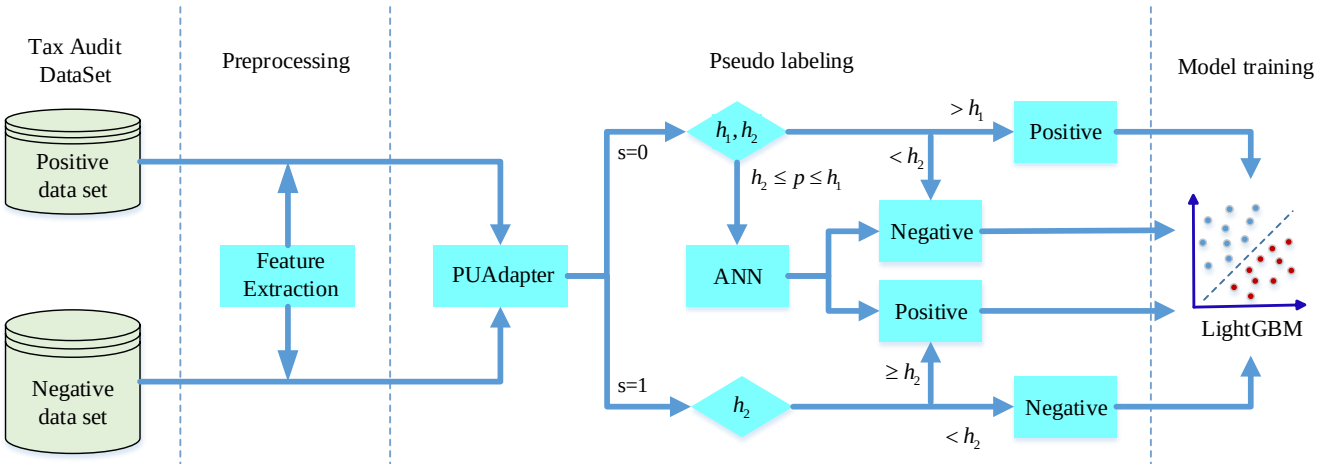


Fig. 1. The framework of the proposed TEDM-PU

In the pseudo labeling stage, the PUAdapter is used to predict the probability that the data belongs to the positive sample. Since the number of positive samples is very small, the effect of the PUAdapter will decrease. To ensure the accuracy of the model in the case of a few positive samples, we divide the data into reliable positive samples, reliable negative samples, and unreliable data according to the probability values. The unreliable samples include data that cannot be clearly distinguished by the PUAdapter, and ANN is used to give pseudo labels for unreliable samples. Finally, each sample has a pseudo label.

In the model training stage, we use the source tax data and pseudo-labels to generate a tax evasion detection model by applying LightGBM classifier.

B. The TEDM-PU Specific Process

1) Preprocessing Stage:

There are two categories of taxpayer features: static information and dynamic information. Static information indicates the properties of the taxpayer, such as the registered capital, number of employees, and age of the legal representative. Dynamic information includes time-series data in the process of interacting with other taxpayers, such as transaction amounts, transaction taxes, and billing days.

In general, tax databases have many features in China. However, not all these features are helpful for evasion detection. In fact, few features would be helpful for tax evasion detection. Therefore, we adopt a random forest classifier to calculate the feature importance based on source data to select the most useful information from the feature space. In the random forest, we use a Gini index [38] to present the importance of features in each decision tree:

$$\begin{aligned}\text{Gini}\,(D, q_j) = &\left(1 - \sum_{i=1}^{n_c} p_{D_i}^2\right) - \\ &\sum_{v=1}^{V} \frac{|D^v|}{|D|}\left(1 - \sum_{i=1}^{n_c} p_{D_i^v}^2\right)\end{aligned} \tag{1}$$

where n_c denotes the number of classes, V denotes the set of all possible values for the feature q_j and p_{D_i} is the ratio of the j^{th} class in data D. The Gini index reflects the probability of an inconsistent category when selecting two samples from the dataset randomly. Then, we calculate the Gini Importance (GI) for each feature q_j as follows:

$$GI = \sum_{i=1}^{n} \frac{C(i)}{|D|}\,\text{Gini}\,(q_j) \tag{2}$$

where n denotes the number of times a feature splits nodes in the random forest and $C(i)$ denotes the number of sample q_j splits. Finally, we select 200 features with the highest GI for tax evasion detection.

2) Pseudo labeling stage:

To train the tax evasion detection model, we need to identify positive and negative samples from positive and unlabeled data. We use the PUAdapter [11] to calculate the probability that each sample belongs to a positive samples.

According to the SCAR hypothesis

$$p(s = 1|x, y = 1) = p(s = 1|y = 1) \tag{3}$$

Then

$$p(y = 1|x) = p(s = 1|x)/c \tag{4}$$

Where $c = p(s = 1|y = 1)$, we can prove as follows:

$$\begin{aligned}p(s = 1|x) &= p(y = 1 \wedge s = 1|x) \\ &= p(y = 1|x)p(s = 1|y = 1, x) \\ &= p(y = 1|x)p(s = 1|y = 1)\end{aligned} \tag{5}$$

The value of the constant $c = p(s = 1|y = 1)$ can be estimated as $c = \frac{1}{n}\sum_{x\in P} g(x)$, where n is the cardinality of P, $g(x) = p(s = 1|x)$. We can prove as follows:

$$\begin{aligned}g(x) =& p(s = 1|x) \\ =& p(s = 1|x, y = 1)p(y = 1|x) \\ &+ p(s = 1|x, y = 0)p(y = 0|x) \\ =& p(s = 1|x, y = 1) \cdot 1 + 0 \cdot 0 \text{ since } x \in P \\ =& p(s = 1|y = 1)\end{aligned} \tag{6}$$

Then there is a $p(y = 1|x)$ for each instance x. For unlabeled samples, if $p(y = 1|x) > h_1$, it is marked as a positive sample. If $p(y = 1|x) < h_2$, it is marked as a reliable negative sample. When the number of positive samples is very small, the accuracy of the PUAdapter will decrease. To ensure the accuracy of the TEDM-PU when there are very few positive examples, if $h_2 \leq p(y = 1|x_i) \leq h_1$, the data cannot be clearly classified. Thus, the data are marked as unreliable. For positive samples, if $p(y = 1|x) < h_2$, it is considered to be a mislabeled sample. Thus it is marked as a negative sample. The positive samples that have been labeled have a higher degree of confidence, so the labels of the other positive samples are unchanged. Apply positive and negative samples training ANN to provide pseudo labels for unreliable samples.

3) Model training stage:

Through the above steps, each sample has a pseudo label, which can be used to train the tax evasion detection model. Using the source tax data and pseudo tags from the above steps, LightGBM is applied to train the tax evasion detection model.

C. The TEDM-PU Procedure

Algorithm 1 TEDM-PU

Input: positive data P; unlabeled data U; test data Z; threshold h_1, h_2.

1: Feature Extraction

2: caluculate c in eq.(6)

$$c = \frac{1}{n} \sum_{x \in P} p(s = 1|x)$$

3: caluculate $p(y = 1|x)$ in eq.(4)

$$p(y = 1|x) = p(s = 1|x)/c$$

4:Pseudo labeling

$$y_i = \begin{cases} 1, & p(y = 1|x_i) > h_1, s = 0 \\ -1, & p(y = 1|x_i) < h_2, s = 0 \\ ANN(x_i), & h_2 \leq p(y = 1|x_i) \leq h_1, s = 0 \\ 1, & p(y = 1|x_i) \geq h_2, s = 1 \\ -1, & p(y = 1|x_i) < h_2, s = 1 \end{cases}$$

5:Model training

$$f = LightGBM(X, Y)$$

Output: hypothesis of test data y

$$y = \begin{cases} 1, & f(x) \geq \frac{1}{2} \\ -1, & f(x) < \frac{1}{2} \end{cases}$$

The procedure of the TEDM-PU is given in Algorithm 1. We apply the PUAdapter to calculate the probability value $p(y = 1|x)$ for each instance that belongs to the positive sample. For unlabeled samples, if $p(y = 1|x) > h_1$, it is marked as a positive sample. If $p(y = 1|x) < h_2$, it is marked as a negative sample. If $h_2 \leq p(y = 1|x_i) \leq h_1$, it is marked as an unreliable sample. For positive samples, if $p(y = 1|x) < h_2$, it is considered to be a mislabeled sample. Thus it is marked as a negative sample. The labels of other positive samples remain unchanged. ANN is applied to positive and negative samples for training to give pseudo labels for unreliable samples. After the above steps, each sample has a pseudo label. Applying LightGBM to train tax evasion detection models using source data and pseudo labels.

V. PERFORMANCE EVALUATION

To evaluate the performance of the TEDM-PU, we first introduce our experimental design. Then we investigate the following research questions:

Question 1: Can our method achieves higher classification accuracy than state-of-the-art approaches? For this, we compared the classification accuracy in thirteen different scenarios using both the proposed and state-of-the-art methods to prove its effectiveness.

Question 2: Can our method be adapted to a small amount of tax evaders and a large amount of unlabeled data? For this purpose, we evaluate the TEDM-PU using a reduced number of tax evaders and compare it with state-of-the-art methods.

Question 3: Can our method explain the detection results and provide evidence of tax evasion? For this purpose, we visualize result of our method to assess the interpretability of the TEDM-PU.

A. Datasets and Metrics

There are no public standard tax audits. We obtained tax data from tax authorities in China to verify our methods. We collected tax information of 20,444 taxpayers in the industrial categories of wholesale and retail. The taxpayers were divided into tax evaders and those without any labels. Each taxpayer has two categories of data: static data and dynamic data. Static data indicates the inherent attributes of taxpayers, including legal representative information (i.e., age, gender, and region), company size, registered capital, and other company-related indicators. The dynamic data comprises a series of trading and declaration data with time-varying properties, such as transaction amounts, transaction taxes, and billing days. For convenience, we named the positive example as P and the unlabeled sample as U. We selected different numbers of positive samples for multiple sets of experiments, as shown in Table II.

In Table II, $|P|$ denotes the amount positive sample used for training, $|U|$ is the amount of unlabeled sample, and $|Z|$ is the size of the test set.

TABLE II
THIRTEEN EXPERIMENT SCENARIOS

Scenarios	$\|P\|$	$\|U\|$	$\|Z\|$
E_1	2000	12310	6134
E_2	1800	12510	6134
E_3	1600	12710	6134
E_4	1400	12910	6134
E_5	1200	13110	6134
E_6	1000	13310	6134
E_7	800	13510	6134
E_8	600	13710	6134
E_9	500	13810	6134
E_{10}	400	13910	6134
E_{11}	300	14010	6134
E_{12}	200	14110	6134
E_{13}	100	14210	6134

The metrics used in our experiments are shown in Table III.

TABLE III
THIRTEEN EXPERIMENT SCENARIOS

Term	Abbr	Definition
True Positive	TP	# of correctly classified tax evaders.
True Negative	TN	# of correctly classified non-tax evaders.
False Negative	FN	# of incorrectly classified non-tax evaders.
False Positive	FP	# of incorrectly classified tax evaders.
Error Rate	ER	$(\mathrm{FN}+\mathrm{FP})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP})$
Precision	P	$\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
Recall	R	$\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$
F-measure	F1	$2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$
ROC Area	AUC	Area under ROC curve

B. Comparison Methods

To verify the performance of the TEDM-PU, we use machine learning and PU learning methods as the comparison method in the experiments. Machine learning methods include RandomForest and SVM. PU learning methods include Bagging SVM [39],Two-Step (NB-SVM) [40] and PUAdapter [11]. Randomforest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. SVM is a discriminative classifier formally defined by a separating hyperplane. A two-step approach is a solution for PU learning problem that consists of tow steps: (1) identifying a set of reliable negative examples, and (2) building a classifier iteratively. PUAdapter is a class prior incorporation positive-unlabeled learning method. Bagging SVM is a solution for PU learning problem based on bootstrap aggregating (bagging) techniques: the algorithm iteratively trains many binary classifiers to discriminate the known positive examples from random subsamples of the unlabeled set and averages their predictions.

C. Experimental Results

1) Effectiveness of the TEDM-PU:

We evaluated the effectiveness of the different methods in different scenarios using three metrics. The evaluation tables are shown in Tables IV to VI.

Table IV shows the error rates of the different methods. For all thirteen scenarios, the TEDM-PU performs better in terms of error rate compared with the traditional machine learning methods and PU learning methods. The error rates of the TEDM-PU are on average 12.7% lower than other methods.

Table V shows the F1 scores of different methods. The TEDM-PU achieves the best performance in all thirteen scenarios. The F1 score of the TEDM-PU are greater than 0.9 in cases where the number of positive samples is greater than 200. Notably, PU learning methods and TEDM-PU maintain excellent performance when the traditional machine learning algorithm performs poorly as the amount of positive samples decreases, which demonstrates the datasets that contains only positive and unlabeled samples cannot be directly classified by traditional machine learning methods.

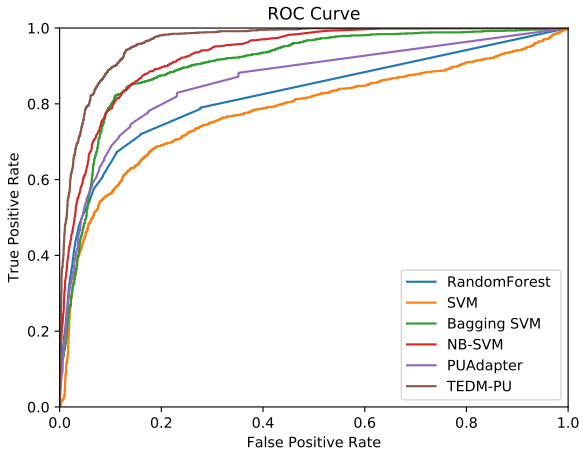


Fig. 2. The ROC Curve of different methods

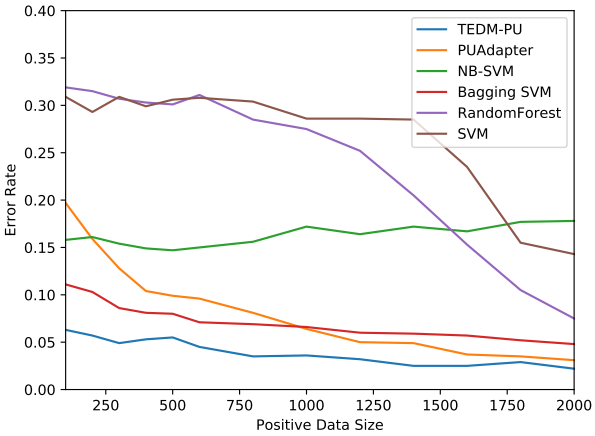


Fig. 3. The error rate of the TEDM-PU with different positive data sizes

As shown in Table VI, the TEDM-PU outperforms other methods and shows outstanding AUC scores, given that the pseudo label acquisition and boosting classifier of the TEDM-PU cause performance gains.

We focus on the ROC curves in E_{13} for different methods, as shown in Figure 2. The TEDM-PU always outperforms other methods and shows outstanding detection accuracy.

In conclusion, the TEDM-PU greatly improves the accuracy of tax evasion identification according to various metrics.

2) Stability of the TEDM-PU:

Figure 3 shows the error rates of different methods with different amounts of positive data. We maintained the total amount of training data and gradually changed the size of the positive data instance. The number of positive instances gradually increased from 100 to 2000. TEDM-PU consistently performs better than other methods. As the positive example of the training set increases, the error rates of SVM and RandomForest continue to decrease, because the problem is similar to the simple two-category problem. In the simple

TABLE IV
ERROR RATES OF TAX EVASION DETECTION WITH DIFFERENT METHODS

Scenarios	SVM	RandomForest	Bagging SVM	NB-SVM	PUAdapter	TEDM-PU
E_1	0.143	0.075	0.048	0.178	0.031	**0.022**
E_2	0.155	0.105	0.052	0.177	0.035	**0.029**
E_3	0.235	0.153	0.057	0.167	0.037	**0.025**
E_4	0.285	0.205	0.059	0.172	0.049	**0.025**
E_5	0.286	0.252	0.06	0.164	0.05	**0.032**
E_6	0.286	0.275	0.066	0.172	0.064	**0.036**
E_7	0.304	0.285	0.069	0.156	0.081	**0.035**
E_8	0.308	0.311	0.071	0.15	0.096	**0.045**
E_9	0.306	0.301	0.08	0.147	0.099	**0.055**
E_{10}	0.299	0.303	0.081	0.149	0.104	**0.053**
E_{11}	0.309	0.307	0.086	0.154	0.128	**0.049**
E_{12}	0.293	0.315	0.103	0.161	0.159	**0.057**
E_{13}	0.309	0.319	0.111	0.158	0.197	**0.063**

TABLE V
F1 SCORE OF TAX EVASION DETECTION WITH DIFFERENT METHODS

Scenarios	SVM	RandomForest	Bagging SVM	NB-SVM	PUAdapter	TEDM-PU
E_1	0.715	0.866	0.919	0.774	0.946	**0.963**
E_2	0.667	0.798	0.917	0.776	0.939	**0.953**
E_3	0.395	0.675	0.908	0.784	0.936	**0.958**
E_4	0.162	0.515	0.904	0.782	0.916	**0.961**
E_5	0.112	0.314	0.899	0.787	0.913	**0.947**
E_6	0.114	0.186	0.889	0.781	0.888	**0.941**
E_7	0.043	0.106	0.889	0.796	0.852	**0.943**
E_8	0.035	0.057	0.879	0.803	0.821	**0.926**
E_9	0.032	0.032	0.862	0.804	0.812	**0.915**
E_{10}	0.026	0.031	0.864	0.799	0.81	**0.919**
E_{11}	0.013	0.013	0.855	0.796	0.752	**0.917**
E_{12}	0.012	0.004	0.823	0.781	0.667	**0.902**
E_{13}	0.004	0.004	0.823	0.769	0.582	**0.891**

TABLE VI
AUC SCORE OF TAX EVASION DETECTION WITH DIFFERENT METHODS

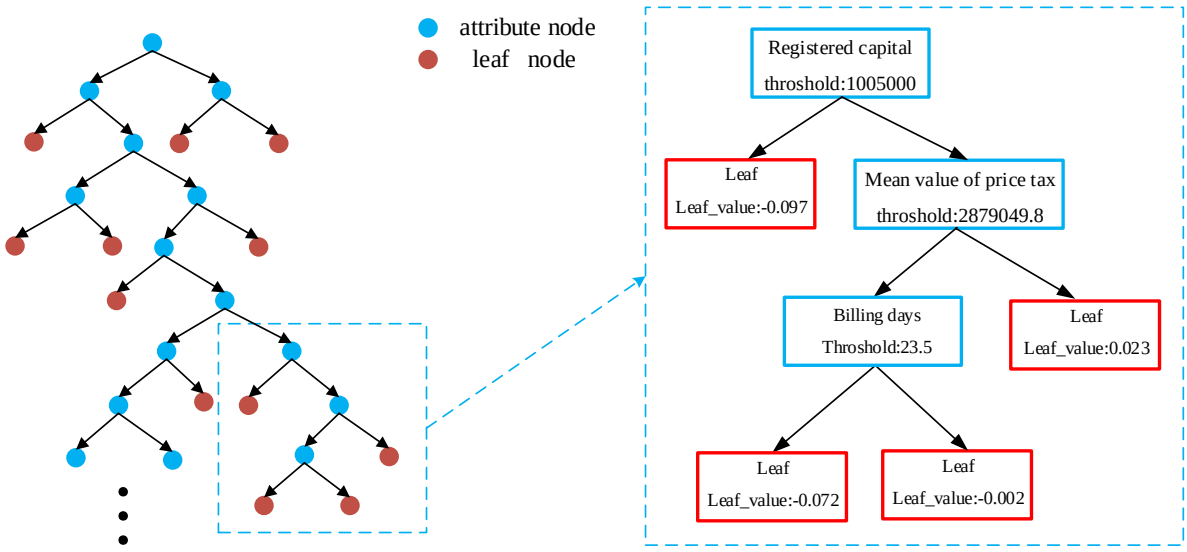
Scenarios	SVM	RandomForest	Bagging SVM	NB-SVM	PUAdapter	TEDM-PU
E_1	0.964	0.932	0.982	0.98	0.991	**0.996**
E_2	0.945	0.912	0.982	0.979	0.99	**0.995**
E_3	0.933	0.9	0.976	0.972	0.988	**0.994**
E_4	0.9	0.909	0.973	0.977	0.985	**0.994**
E_5	0.895	0.889	0.977	0.978	0.983	**0.995**
E_6	0.875	0.892	0.974	0.974	0.979	**0.992**
E_7	0.866	0.886	0.974	0.976	0.975	**0.993**
E_8	0.833	0.87	0.975	0.97	0.96	**0.99**
E_9	0.826	0.867	0.968	0.97	0.958	**0.984**
E_{10}	0.844	0.861	0.968	0.969	0.942	**0.991**
E_{11}	0.828	0.852	0.966	0.97	0.92	**0.989**
E_{12}	0.825	0.825	0.958	0.957	0.894	**0.986**
E_{13}	0.801	0.785	0.942	0.944	0.88	**0.983**

two-category problem, as the positive example increases, the traditional machine learning algorithm can be well classified. When the positive example is gradually reduced, the error rate of the PU learning method is significantly lower than that of SVM and RandomForest. The TEDM-PU is better than Bagging SVM, NB-SVM, and PUAdapter due to the advantages of the pseudo label acquisition and boosting technique. The TEDM-PU can use unlabeled data for tax evasion detection when there are very few positive examples in the training data. Therefore, TEDM-PU has excellent performance and stability in tax evasion detection with only positive and unlabeled instances.

3) Interpretability of the TEDM-PU:

Most machine learning-based testing methods do not clearly explain the tax evasion process when detecting tax evaders. The TEDM-PU has the advantage of interpretability, thus it has the ability to interpret the detection process. Figure 4 shows a visualization of the training results of the TEDM-PU.

Figure 4 consists of two types nodes of different colors. The blue node is an attribute node, indicating the feature name and corresponding threshold for the node to be split. The red node is a leaf node that will give a value to the sample for the next



(a) visualization of training results (b) partial visualization

Fig. 4. Interpretability of the TEDM-PU

iteration. Positive values represent the suspicion of tax evasion, while negative values indicate normal findings. We can obtain the suspicious value of tax evasion by assessing the value of the leaf node passed by the taxpayer. Therefore, if the TEDM-PU detects tax evaders, we can determine which indicators are abnormal. The TEDM-PU can tell which indicators are abnormal using a tree structure. In addition, TEDM-PU can highlight the link between various indicators. It is normal to assess a single indicator, but problems can occur when indicators are grouped together. In practice, the TEDM-PU can help with tax evasion detection as a supplementary audit tool.

D. Further discussion

1) Advantages:

Based on the above evaluation results, we summarize the advantages of our approach below.

Practicality: The proposed approach benefits from the advantages of PU learning and a boosting classifier. The intent of this hybrid solution is to solve the problem of training data with only a small amount of tax evaders and a large number of unlabeled tax evaders during tax evasion detection. The performance test shows that our approach achieved the best performance for all metrics.

Stability: The proposed method can be applied to detect tax evasion with a small amount of tax evader and a large amount of unlabeled taxpayer. Our performance evaluation different amount of tax evader, and the results imply that our approach can converge to the best performance with a small amount of tax evader.

Accuracy: Based on our evaluation, the proposed approach can achieve higher tax evasion detection accuracy than state-of-the-art methods using different amount of tax evader. In addition, our method converges to a low error rate quickly with a small amount of tax evader.

Interpretability: The proposed approach is interpretable. When a tax evader is detected based on some features of the taxpayer, it can provide a detailed tax evasion detection trail, which explains the reasons for tax evasion. Auditor can deduce detailed information from the visualization of the model.

2) Disadvantages:

There are some disadvantages in our approach. We must further improve the following aspects. First, we must collect a small amount of tax evaders to construct the tax detection model. Second, to improve the accuracy of the approach, we adopted the pseudo labeling technique but it increases the computation time.

VI. CONCLUSION

This paper proposed the TEDM-PU, a novel tax evasion detection method. It integrates PU learning and an interpretable classifier to train a tax detection model and induce interpretability in the detection model. First, by combining feature-based and class prior incorporation PU learning techniques, the probability value for each instance that belongs to the positive sample is calculated. Second, a pseudo tag is calculated for each sample to augment the learning data . Finally, to offer a clear explanation of tax evasion, LightGBM is adopted to build an interpretable and accurate tax evasion detection model. The TEDM-PU outperforms state-of-the-art methods with higher accuracy. In addition, it provides a good explanation of tax evasion behavior.

With regard to future studies, we will improve the design and seek a new method to optimize the speed of the process. Moreover, we aim to achieve unsupervised tax evasion detection in conjunction with tax expert knowledge.

ACKNOWLEDGMENT

This research was partially supported by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with Grant No. 2018YFB1004500, the MOE Innovation Research Team No. IRT17R86, the National Science Foundation of China under Grant Nos. 61721002 and 61532015, the Project of China Knowledge Centre for Engineering Science and Technology, and the consulting research project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China.

REFERENCES

[1] J. Ruan, Z. Yan, B. Dong, Q. Zheng, and B. Qian, "Identifying suspicious groups of affiliated-transaction-based tax evasion in big data," Information Sciences, vol. 477, pp. 508-532, Mar. 2019.

[2] F. Tian et al., "Mining Suspicious Tax Evasion Groups in Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 10, pp. 2651-2664, Oct. 2016.

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[5] J. Bekker and J. Davis, "Learning From Positive and Unlabeled Data: A Survey," arXiv:1811.04820 [cs, stat], Nov. 2018.

[6] M. Claesen, F. De Smet, P. Gillard, C. Mathieu, and B. De Moor, "Building Classifiers to Predict the Start of Glucose-Lowering Pharmacotherapy Using Belgian Health Expenditure Data," arXiv:1504.07389 [cs, stat], Apr. 2015.

[7] K. Zupanc and J. Davis, "Estimating rule quality for knowledge base completion with the relationship between coverage assumption," in Proceedings of the Web Conference 2018, 20180401, pp. 1-9.

[8] A. Kaboutari, J. Bagherzadeh, and F. Kheradmand, "An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification," IJCATR, vol. 3, no. 9, pp. 592-594, Sep. 2014.

[9] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," in Proceedings of the 18th International Joint Conference on Artificial Intelligence, San Francisco, CA, USA, 2003, pp. 587-592.

[10] X. Zhu, Z. Yan, J. Ruan, Q. Zheng, and B. Dong, "IRTED-TL: An Inter-Region Tax Evasion Detection Method Based on Transfer Learning," 2018, pp. 1224-1235.

[11] C. Elkan and K. Noto, "Learning Classifiers from Only Positive and Unlabeled Data," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2008, pp. 213-220.

[12] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3146-3154.

[13] Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, vol. 61, pp. 85-117, Jan. 2015.

[14] R.-S. Wu, C. S. Ou, H. Lin, S.-I. Chang, and D. C. Yen, "Using data mining technique to enhance tax evasion detection performance," Expert Systems with Applications, vol. 39, no. 10, pp. 8769-8777, Aug. 2012.

[15] Denny, G. J. Williams, and P. Christen, "Exploratory Multilevel Hot Spot Analysis: Australian Taxation Office Case Study," in Proceedings of the Sixth Australasian Conference on Data Mining and Analytics - Volume 70, Darlinghurst, Australia, Australia, 2007, pp. 77-84.

[16] Z. Assylbekov, I. Melnykov, R. Bekishev, A. Baltabayeva, D. Bissengaliyeva, and E. Mamlin, "Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan," 2016, pp. 37-49.

[17] X. Liu, D. Pan, and S. Chen, "Application of Hierarchical Clustering in Tax Inspection Case-Selecting," in 2010 International Conference on Computational Intelligence and Software Engineering, 2010, pp. 1-4.

[18] P. Castellon Gonzalez and J. D. Velasquez, "Characterization and detection of taxpayers with false invoices using data mining techniques," Expert Systems with Applications, vol. 40, no. 5, pp. 1427-1436, Apr. 2013.

[19] Y.-S. Chen and C.-H. Cheng, "A Delphi-based rough sets fusion model for extracting payment rules of vehicle license tax in the government sector," Expert Systems with Applications, vol. 37, no. 3, pp. 2161-2174, Mar. 2010.

[20] E. Junque de Fortuny, M. Stankova, J. Moeyersoms, B. Minnaert, F. Provost, and D. Martens, "Corporate Residence Fraud Detection," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2014, pp. 1650-1659.

[21] D. DeBarr and Z. Eyler-Walker, "Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters," SIGKDD Explor. Newsl., vol. 8, no. 1, pp. 11-16, Jun. 2006.

[22] M. Gupta and V. Nagadevara, "Audit Selection Strategy for Improving Tax Compliance - Application of Data Mining Techniques," 2007.

[23] J. A. Noguera, F. J. M. Quesada, E. Tapia, and T. Llacer, "Tax Compliance, Rational Choice, and Social Influence: An Agent-Based Model," Revue francaise de sociologie, vol. Vol. 55, no. 4, pp. 765-804, 2014.

[24] T. Llacer, F. J. Miguel, J. A. Noguera, and E. Tapia, "AN AGENT-BASED MODEL OF TAX COMPLIANCE: AN APPLICATION TO THE SPANISH CASE," Advances in Complex Systems (ACS), vol. 16, no. 04n05, pp. 1-33, 2013.

[25] L. Antunes, J. Balsa, and H. Coelho, "Agents that collude to evade taxes," in Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems - AAMAS '07, Honolulu, Hawaii, 2007, p. 1.

[26] N. Abe et al., "Optimizing Debt Collections Using Constrained Reinforcement Learning," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2010, pp. 75-84.

[27] N. D. Goumagias, D. Hristu-Varsakelis, and A. Saraidaris, "A decision support model for tax revenue collection in Greece," Decision Support Systems, vol. 53, no. 1, pp. 76-96, Apr. 2012.

[28] D. He and D. Wang, "Robust Biometrics-Based Authentication Scheme for Multiserver Environment," IEEE Systems Journal, vol. 9, no. 3, pp. 816-823, Sep. 2015.

[29] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in In: Intl. Conf. on Data Mining, 2003, pp. 179-188.

[30] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," in Proceedings of the Nineteenth International Conference on Machine Learning, San Francisco, CA, USA, 2002, pp. 387-394.

[31] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," in Proceedings of the 18th International Joint Conference on Artificial Intelligence, San Francisco, CA, USA, 2003, pp. 587-592.

[32] M. Claesen, F. De Smet, J. A. K. Suykens, and B. De Moor, "A robust ensemble approach to learn from positive and unlabeled data using SVM base models," Neurocomputing, vol. 160, pp. 73-84, Jul. 2015.

[33] Y. Zhang, X. Ju, and Y. Tian, "Nonparallel hyperplane support vector machine for PU learning," in 2014 10th International Conference on Natural Computation (ICNC), 2014, pp. 703-708.

[34] S. Sellamanickam, P. Garg, and S. K. Selvaraj, "A Pairwise Ranking Based Approach to Learning with Positive and Unlabeled Examples," in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2011, pp. 663-672.

[35] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R Leathwick, "Presence-Only Data and the EM Algorithm," Biometrics, vol. 65, pp. 554-63, Sep. 2008.

[36] F. Denis, R. Gilleron, and F. Letouzey, "Learning from positive and unlabeled examples," Theoretical Computer Science, vol. 348, no. 1, pp. 70-83, Dec. 2005.

[37] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," Forest, vol. 23, Nov. 2001.

[38] L. Breiman, Classification and Regression Trees. Routledge, 2017.

[39] F. Mordelet and J.-P. Vert, "A bagging SVM to learn from positive and unlabeled examples," Pattern Recognition Letters, vol. 37, pp. 201-209, Feb. 2014.

[40] A. Kaboutari, J. Bagherzadeh, and F. Kheradmand, "An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification," International Journal of Computer Applications Technology and Research, vol. 3, pp. 592-594, Sep. 2014.

[41] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-Unlabeled Learning with Non-Negative Risk Estimator," in Advances

in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 1675-1685.

[42] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," Third IEEE International Conference on Data Mining, pp. 179-186, 2003.