Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence–Review

A Survey of Tax Risk Detection Using Data Mining Techniques

Qinghua Zheng [a,b], Yiming Xu [a,b], Huixiang Liu [a,b], Bin Shi [a,b,*], Jiaxiang Wang [a,b], Bo Dong [b,c]

[a] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[b] Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[c] School of Distance Education, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Tax risk detection
Data mining
Knowledge guide
Informatization
Intellectualization

ABSTRACT

Tax risk behavior causes serious loss of fiscal revenue, damages the country's public infrastructure, and disturbs the market economic order of fair competition. In recent years, tax risk detection, driven by information technology such as data mining and artificial intelligence, has received extensive attention. To promote the high-quality development of tax risk detection methods, this paper provides the first comprehensive overview and summary of existing tax risk detection methods worldwide. More specifically, it first discusses the causes and negative impacts of tax risk behaviors, along with the development of tax risk detection. It then focuses on data-mining-based tax risk detection methods utilized around the world. Based on the different principles employed by the algorithms, existing risk detection methods can be divided into two categories: relationship-based and non-relationship-based. A total of 14 risk detection methods are identified, and each method is thoroughly explored and analyzed. Finally, four major technical bottlenecks of current data-driven tax risk detection methods are analyzed and discussed, including the difficulty of integrating and using fiscal and tax fragmented knowledge, unexplainable risk detection results, the high cost of risk detection algorithms, and the reliance of existing algorithms on labeled information. After investigating these issues, it is concluded that knowledge-guided and data-driven big data knowledge engineering will be the development trend in the field of tax risk in the future; that is, the gradual transition of tax risk detection from informatization to intelligence is the future development direction.

1. Introduction

Tax revenue is a country's most important source of revenue. In 2021, the Chinese tax authorities determined that tax revenues (with export tax rebates deducted) accounted for 76.3% of national general public budget revenue [1]. However, the biggest challenges related to tax governance in countries around the world remain the issues of tax evasion, fraud, and other tax risks. Tax risk behaviors such as tax evasion and fraud cause serious tax losses, erode the tax base, and damage national tax security. At the same time, such behaviors disrupt the market economic order of fair competition and give rise to unfair competition. The average total tax revenue loss in the United States from 2011 to 2013 has been estimated at 441 billion USD per year [2]. In the European Union, according to data provided in 2015, the average share of tax loss is 18.3%; in Switzerland, the country with the lowest share, tax loss is still equivalent to 6.9% of the gross domestic product (GDP) [3]. The World Bank estimates that 54% of companies in 135 developing countries do not declare all income taxes to the tax authorities [4]. Combating and rectifying tax risk behaviors are therefore crucial requirements for regulating market order and maintaining fairness and justice, as well as being important measures to improve national governance capability.

The key in combating and rectifying tax risk behaviors is accurate tax risk detection [5]. However, this task is highly complex and difficult, due to the following aspects:

(1) **Complex, high-dimensional, and massive data.** Tax scenarios contain trillions of data points from different sources (e.g., industry and commerce, taxation, social security, customs, etc.), as well as different types of data (e.g., regulations and policies, statements, invoices, contracts, and transaction documents).

* Corresponding author at: School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.
E-mail address: shibin@xjtu.edu.cn (B. Shi).

https://doi.org/10.1016/j.eng.2023.07.014

Please cite this article as: Q. Zheng, Y. Xu, H. Liu et al., A Survey of Tax Risk Detection Using Data Mining Techniques, Engineering, https://doi.org/10.1016/j.eng.2023.07.014

(2) **Hidden tax evasion mode.** In the process of the constant game between tax inspectors and tax evaders, tax evasion and fraud behavior have developed to encompass gangs, specialization, and concealment, with examples including registering or purchasing a large number of shell companies, layer-upon-layer covering, lengthening the timing chain of crime, and wantonly committing crimes across regions, industries, and even borders to evade supervision and crackdown.

In light of the issues described above, traditional tax risk detection methods such as manual case selection, reporting case selection, and rule-based case selection are greatly limited due to their high dependence on manpower and the knowledge of financial and tax experts. As a result, hidden and complicated tax-related crimes are difficult to identify. The question of how to efficiently mine hidden tax risk clues from massive multi-source heterogeneous tax-related data has become an urgent one for tax authorities and scientific researchers alike.

In recent years, tax risk detection methods based on artificial intelligence and data mining have attracted extensive attention in many countries and regions, and numerous excellent methods have emerged. Methods of this kind can mine the internal relations between and laws of data within massive structured and unstructured data, thereby providing an effective solution to the above bottleneck problems and a new paradigm for tax risk detection. Such methods can more easily mine deeper knowledge from the complex structural information and rich semantic information contained in the tax transaction network, thereby improving the accuracy of tax risk detection.

In this survey, we review existing tax risk detection methods. This survey aims to introduce the relevant background knowledge and development process of tax risk detection, comprehensively sort existing methods, summarize the main problems encountered in tax risk detection today, and explore future research directions in the field, with the goal of facilitating tax risk detection research and the development of more powerful methods to combat tax risk behaviors effectively. The main contributions of this survey can be summarized as follows:

(1) To the best of our knowledge, we are the first to systematically review research progress and development trends in tax risk detection worldwide. We hope to promote the exploration of more powerful tax risk detection methods, and thereby reduce national tax losses.
(2) We introduce relevant background knowledge related to tax risk detection, including the causes and negative impacts of tax risk behaviors, along with the development process of tax risk detection. In addition, we provide a formal definition of tax risk detection and introduce details of the input data.
(3) We screen 89 tax risk detection research works from around the world based on relevance and importance, comprehensively sort the research on tax risk detection, divide the existing methods into two categories (relationship-based risk detection methods and non-relationship-based risk detection methods), and then list and introduce the 14 identified methods. Furthermore, we summarize the advantages and disadvantages of each method, which is critical for practical applications.
(4) We summarize the main problems faced in current tax risk detection practice and suggest a list of future research directions that must be urgently pursued to advance tax risk detection from informatization to intelligence.

This survey is organized as follows. In Section 2, we introduce the background related to tax risk detection. We then categorize existing tax risk detection algorithms and formally define the tax risk detection problem. In Sections 3 and 4, we comprehensively review existing tax risk detection methods. Future research directions in the field of tax risk detection are then explored in Section 5. Finally, we summarize the paper in Section 6.

2. Background

In this section, we first introduce the relevant background knowledge on tax risk detection, including the causes and negative impacts of tax risk behaviors. Next, we discuss the nature of the input data in tax scenarios. We then present the development process and types of tax risk detection to provide a clear overview of the field. Finally, we present a problem formulation for tax risk detection.

2.1. Causes and negative impacts of tax risk behaviors

Tax risk behavior is essentially an illegal activity driven by profit. Taxpayers can make high profits through tax fraud or by deliberately avoiding the payment of their true tax obligations. However, tax risk behaviors not only seriously erode the tax base but also indirectly affect the competitiveness of lawful and honest taxpayers, thereby disrupting the market economic order of fair competition.

2.2. Nature of tax data

The most important part of tax risk detection is tax-related data. Risk detection models mine risk information from massive data to achieve risk management. Since the quantity of tax payable is related to the company's whole production and operation, tax-related data naturally include the information of taxpayers and their operations. Depending on the nature of tax data, the input data used in tax risk detection can be classified into contextual attributes and behavioral attributes.

2.2.1. Contextual attributes

Contextual attributes are used to determine the context and inherent properties of an entity object. We selected some important contextual attributes related to tax risk detection, which are listed in Table 1. These contextual attributes consist of three categories: registration information, financial statements, and business information. Accordingly, the attribute types can be divided into three categories: number, enumeration (enum), and text.

2.2.2. Behavioral attributes

Behavioral attributes define the non-contextual characteristics of entity objects and describe the various types of relationships between these entity objects, such as up-/down-stream information and transaction value. We selected some important behavioral

Table 1
Contextual attributes.

Category	Attributes	Type
Registration information	Registration type	Enum
	Taxpayer current status	Enum
	Industry code	Enum
	Registration address	Text
	Business scope	Text
	Registration capital	Number
Financial statements	Total investment	Number
	Annual sales	Number
	Profit margin	Number
	Total investment	Number
Business information	Number of employees	Number
	Age of legal person	Number
	Corporate credit rating	Enum
	Business items	Text

attributes related to tax risk detection, which are listed in Table 2. The behavioral attributes can be divided into two categories: statistical information and up-/down-stream information.

2.3. The development of tax risk detection

Tax risk detection is the practice of effectively dealing with tax risks; it is the concentrated embodiment of the national tax governance level. Tax risk detection has undergone two main stages of evolution: traditional case selection and data mining-based case selection.

(1) **Traditional case selection.** In the early stages, tax risk detection mainly employed the traditional case selection method, which can be further subdivided into report-based case selection methods, manual-based case selection methods, and rule-based case selection methods. Report-based case selection methods rely primarily on reporting information and have strong contingency. Manual-based case selection methods are mainly dependent on the manpower of finance and taxation experts. Notably, with rapid economic development, the total quantity of finance and taxation data has exploded, and it is no longer possible to achieve adequate risk detection for all enterprises using manual inspections. Rule-based case selection methods typically employ expert experience to define rules, and then establish a rule-based reasoning system to identify risk points in fiscal and taxation data. Such rule-based case selection methods are reliant on the experience and knowledge of tax experts in audit work and often encounter problems such as strong subjectivity and lag (i.e., rules cannot be updated in a timely manner), making it difficult to detect new tax risk behaviors.

(2) **Data mining-based case selection.** In order to mitigate the limitations of traditional case selection methods, case selection methods based on data mining have been extensively studied at home and abroad. This approach guides tax audit work through the constant learning of historical data. Due to the numerous advantages of data-mining-based tax risk detection methods in comparison with traditional case selection methods, the former achieve superior performance and have accordingly received widespread attention. The numbers of published papers on data-mining-based tax risk detection methods at home and abroad every year since 1999 are shown in Fig. 1. From the figure, it can be seen that research on data-mining-based tax risk detection is attracting increasing attention from researchers, with the number of papers on the subject published over the past two decades exhibiting an overall increasing trend.

According to the use of the input data, tax risk detection methods based on artificial intelligence and data mining can be divided into two categories: non-relationship-based and relationship-based tax risk detection. Generally speaking, non-relational methods use only the contextual attributes (see Section 2.2.1.) of taxpayers to identify risks, without considering the interaction between taxpayers. In fact, however, there are various types of relationships between different entity objects in a tax scenario. In order to make full use of the rich behavioral attributes between entities in tax scenarios (see Section 2.2.2.), relationship-based risk detection methods have come into being. The existing tax risk detection methods can be further classified into 14 types, as shown in Fig. 2. These 14 types of methods are discussed in more detail in Sections 3 and 4.

Table 2
Behavioral attributes.

Category	Attributes	Type
Statistical information	Average transaction value	Number
	Transaction value variance	Number
	Total transaction value	Number
	Median transaction value	Number
	Minimum transaction value	Number
	Maximum transaction value	Number
	Average tax rate	Number
Up-/down-stream information	Proportion of upstream invoices	Number
	Proportion of downstream invoices	Number

2.3.1. Non-relationship-based tax risk detection

To identify risk individuals, non-relationship-based tax risk detection methods first extract the relevant features of risk individuals, then train the classifier, and finally carry out risk detection. As shown in Fig. 3, these methods use only the contextual attributes (see Section 2.2.1.) of risk individuals; they do not consider the rich interaction information between them—namely, the behavioral attributes.

2.3.2. Relationship-based tax risk detection

As discussed above, a tax scenario contains different kinds of entities, along with the rich interactions between them. However, non-relational methods do not explore the behavioral attributes that describe the interactions between taxpayers. In order to make full use of the rich interactions between entities, relationship-based risk detection methods integrate the characteristics of risk individuals with the relationships and structural characteristics between them in the tax transaction network, in order to identify risks in terms of risk groups. As shown in Fig. 3, these methods use both the contextual attributes of taxpayers and the behavioral attributes (see Section 2.2.2.) shared between them.

Non-relational data mining methods were developed earlier than relational methods, and the research works related to non-relational methods are more numerous. As shown in Fig. 1, since 2006, many scholars have employed non-relational data mining methods to identify tax risks. These approaches only make use of the contextual attributes (see Section 2.2) of tax-related entities and usually employ methods such as feature engineering to analyze the data of tax-paying enterprises; they then select a set of features that can reflect the risk of tax-paying enterprises (e.g., personal characteristics, business characteristics, tax characteristics, bill characteristics, etc.) and subsequently train classifiers based on the selected features (e.g., support vector machines (SVMs), clustering models, etc.). Later, in order to make full use of the interactive relationships between various entities in tax scenarios, relationship-based data mining case selection methods were developed. After 2016, scholars gradually began to use complex relationships in tax networks to identify risks. Due to their use of both contextual attributes and behavioral attributes (see Section 2.2), relation-based methods can more easily mine deeper knowledge from the complex structural information and rich semantic information contained in the tax network, thereby improving the accuracy of risk detection. The trend in tax risk detection techniques is for models to focus on as much information as possible in the tax scenario, rather than solely exploring the contextual features of the company.

2.4. Problem formulation of tax risk detection

In a tax risk detection problem, various objects and their interactions can be modeled as $G(V, E)$, where V is the set of entities in the tax scenario, including multiple entities such as taxpayers, while E is the set of relationships between entities. N is the number of all entities in G, F is the number of dimensions of the contextual features, and the contextual features of all entities are defined as $\boldsymbol{X} \in \mathbb{R}^{N \times F}$. $\boldsymbol{E} \in \mathbb{R}^{N \times N \times P}$ denotes the matrix of behavioral features in

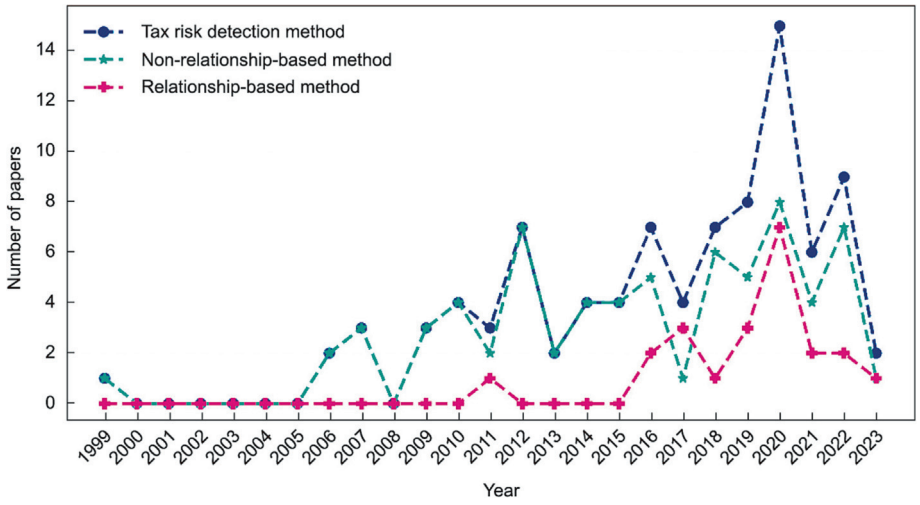

Fig. 1. Research and development trends of tax risk detection.

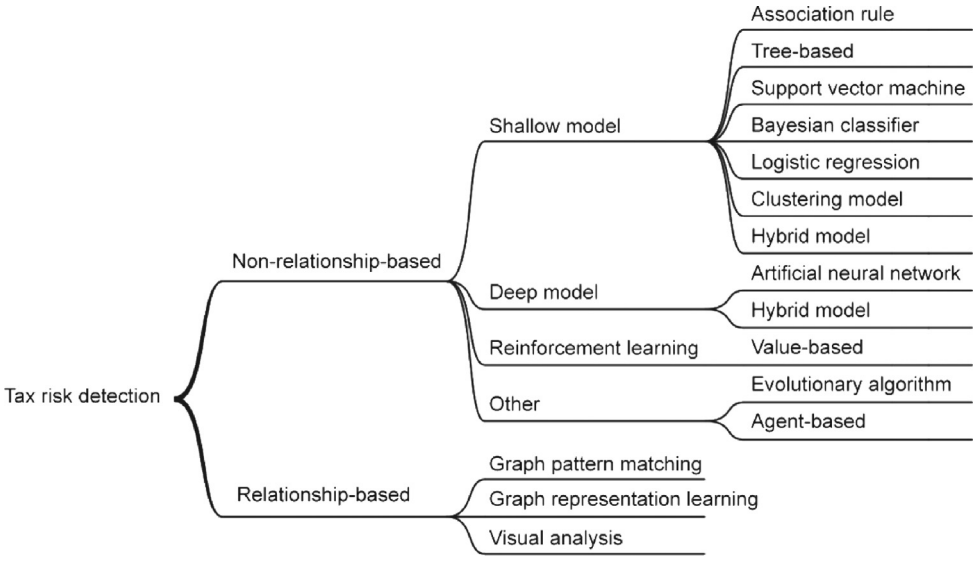

Fig. 2. Classification of tax risk detection methods.

G, P is the dimension of the behavioral features, $\mathbb{R}$ is real number space and $\boldsymbol{E}_{ij\cdot} \in \mathbb{R}^P$ denotes the behavioral features between entity i and entity j. The goal of the tax risk detection model is to learn a mapping function f that can detect risky taxpayer behavior based on contextual features, behavioral topology information, and behavioral features. The formal representation is as follows:

$$\boldsymbol{Y} = f(V, E, \boldsymbol{X}, \boldsymbol{E}) \tag{1}$$

where $\boldsymbol{Y}$ is the risk score vector of all taxpayers, with higher scores implying higher levels of risk.

3. Non-relationship-based tax risk detection methods

3.1. Non-relationship-based shallow models

3.1.1. Association rule

Association rule is a rule-based machine learning method that is one of the most important techniques in the field of data mining. It is used to discover correlations and patterns between certain attributes in massive data [6,7].

Wu et al. [8] used association rules for a value-added tax (VAT) database to discover patterns and relationships between attributes in VAT evasion reports. The researchers developed a screening framework based on specific patterns or rules present in the VAT evasion reports, and then screened cases that did not match the VAT reports for further auditing. This model helps tax auditors perform tax evasion inspections more effectively and improves the hit rate when screening tax evasion cases.

A study by Matos et al. [9] investigated the detection of tax evasion in Brazil. The researchers applied association rules to identify tax fraud patterns and applied two-dimensional (2D) reduction methods (i.e., principal component analysis (PCA) and singular value decomposition (SVD)) to generate a fraud scale ranking taxpayers according to their likelihood of engaging in fraudulent behavior. The results of the study showed that the model can identify tax fraudsters with 80% accuracy.

The association rule algorithm is simple, and its results are easy to understand. However, as the amount of data increases, the computational complexity also increases significantly, due to the higher number of candidate sets [10].

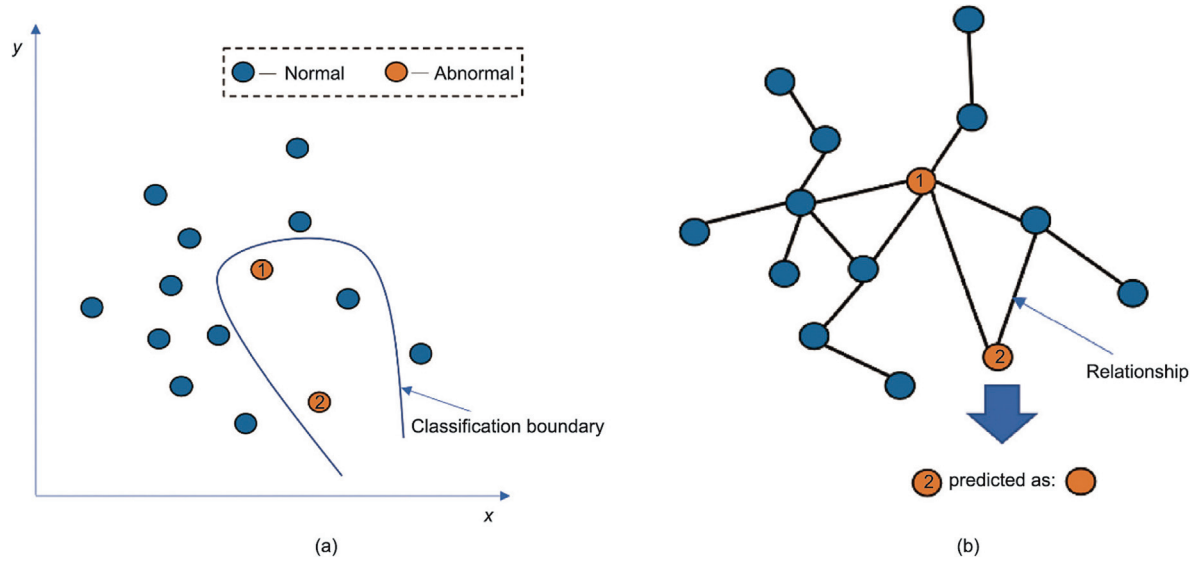

Fig. 3. The difference between non-relationship-based and relationship-based tax risk detection. (a) Non-relationship-based tax risk detection; (b) relationship-based tax risk detection.

3.1.2. Tree-based models

The tree-based model, also known as the decision tree, is one of the best-known machine learning algorithms; it is commonly used in tasks involving statistics and data mining [11,12]. A tree-based model usually contains a root node, several internal nodes, and several leaf nodes. Leaf nodes indicate decision results, other nodes represent judgments about attributes, and each branch represents the output of a judgment.

Bonchi et al. [13] used a case study to illustrate how decision tree-based classification techniques can be used to assist in the design of audit strategies. Such research requires a trade-off between maximizing audit benefits and minimizing costs. The researchers suggested that an appropriate integration of deductive reasoning (e.g., reasoning supported by a logical database) and inductive reasoning (e.g., reasoning supported by decision trees) can provide effective solutions to many problems. The combined use of these two approaches in reasoning is especially effective during the evaluation phase.

Mittal et al. [14] applied a random forest-based classifier to identify fraudulent enterprises in Delhi's VAT system. The classifier uses tax data and reports from tax officials as a training set to predict the likelihood of an enterprise being fraudulent. Experiments showed that tax administrations can recover tens of millions of dollars in losses due to fraud through this method. In addition, the researchers contend that each tax jurisdiction has its own unique characteristics, meaning that the work needs to be highly policy-relevant.

Yao et al. [15] proposed a hybrid method to detect financial fraud. The researchers combined feature selection and classification algorithms to optimize their model. They calculated and analyzed the factors affecting fraudulent behavior, considered the influence of the number of variables on the model, and finally compared the performance of five machine learning algorithms through experiments. The results indicated that the random forest algorithm performs well, is suitable for analyzing and processing high-dimensional data, and can effectively avoid overfitting problems.

Wu et al. [16] proposed a tax evasion detection method based on random forest. This study was aimed at the automobile sales industry; it employed operating data, taxpayer attributes, and operating characteristics as input, and applied a random forest model to identify taxpayers exhibiting tax evasion behaviors. In the process of constructing the random forest, *k*-fold cross-validation was used to select the optimal number of decision trees. The experiment also compared random forest with AdaBoost, logistic regression (LR), and other algorithms. The results showed that random forest had the highest accuracy in this task and could precisely detect tax evasion.

An et al. [17] proposed a method for constructing a classification model that achieved excellent performance in the task of identifying financial statement fraud; they further provided decision rules that could be used to explain the classification results. The proposed model, which is largely an improvement on the random forest algorithm, demonstrates good classification performance and can obtain classification rules. This study addressed the problem of class imbalance: The researchers divided the imbalanced dataset into multiple balanced sub-datasets, randomly selected features in each sub-dataset, and trained an appropriate number of classification and regression tree (CART) decision trees. Finally, the model with the best accuracy was screened using the test data; this optimal model was called the modified random forest (MRF) model. The MRF model was then used to detect financial statement fraud.

Considering that the combination of expert experience and machine learning was a feasible method to control tax risk, Ji et al. [18] selected certain commercial enterprises as experimental cases, constructed a random forest algorithm, and established a detection model to assess the risk of illegally issuing false invoices. An analysis of the experimental results revealed that the model has high accuracy and can be used as a reference for tax reviewers. Andrade et al. [19] presented a machine learning-based system for detecting tax fraud in the state of Espírito Santo (Brazil) by classifying companies' financial data. Four different classifiers—*k*-nearest neighbors, a random forest, an SVM, and a neural network—were trained and tested, with the random forest achieving the best macro-averaged F1-score of 92.98%. The system was shown to reduce manual workload by 81%, with a loss of 15% of fraudulent companies. Future work includes incorporating new features, addressing the data imbalance problem, and transferring learning across geographic regions.

Xavier et al. [20] proposed a solution for identifying tax evading companies using open and public data, achieving an accuracy of over 98% with a random forest model. The researchers emphasized the feasibility of using public data to tackle tax evasion and

demonstrated the potential of using neural networks for heterogeneous and relational graphs. The reported system is already in use by tax auditors and delegates, indicating its practical application. Overall, this paper offers a promising approach for tax evasion identification using open data and artificial intelligence.

Overall, tree-based models perform well, can handle high-dimensional samples with missing attributes, and produce results that are easy to understand and interpret. However, tree-based models are also prone to overfitting, although random forests and tree pruning can alleviate overfitting [21]. Finally, it is difficult for tree models to support online learning, and it is often necessary to rebuild the decision tree in these contexts.

3.1.3. Support vector machine

An SVM is a classification model—more specifically, a binary linear classifier with the largest interval defined on the feature space [22,23]. Its learning strategy is interval maximization, which can be formalized as a convex quadratic optimization problem. In addition to addressing linear classification tasks, an SVM can use kernel tricks to implicitly map inputs into a high-dimensional space, effectively performing nonlinear classification.

Wang and Li [24] introduced the use of an SVM in tax fraud detection. Their study regarded the task as a binary category problem with "trusted" and "untrusted" categories. In the experiment, the VAT payments of 61 retail enterprises in Qingdao were tested, and the SVM was used to determine the category of enterprises. The final accuracy rate reached 87.10%.

Liu et al. [25] combined the advantages of rough set theory and an SVM to propose a new tax assessment model. First, a tax assessment index system was established, after which the attributes of the assessment index were reduced using rough set theory. The reduced index was then regarded as the input of the SVM, thus forming the overall framework of the model. It is worth noting that the kernel function of the SVM in this study adopted a radial basis function (RBF) kernel function. The experimental results showed that the model performs well; moreover, in combination with rough set theory, it exhibited higher accuracy and a shorter training time than an SVM alone.

Xia and Li [26] proposed a tax detection method that combined an SVM with a self-organizing map (SOM). First, the SVM was used to classify taxpayers. In the SVM training process, a genetic algorithm was also introduced to take advantage of its excellent search characteristics. The classified taxpayer information was then used as input, and the SOM was employed to cluster the taxpayers in order to facilitate further inspection by auditors.

Junqué de Fortuny et al. [27] used structured and fine-grained invoice data for fraud detection and constructed a tax fraud detection approach with the advantages of being efficient and easy to use. They focused on the transaction invoice data between foreign companies in Belgium. The analysis showed that the input data has the characteristics of high dimensionality and high sparsity. Thus, two different methods could be used to analyze the input data: one was an SVM or naive Bayes, and another was relational learning on graph representations. After considering the trade-off between performance and interpretability, a linear SVM model was finally selected.

Rad et al. [28] analyzed, designed, and implemented a system for predicting high-risk taxpayers, which was used to predict the behavior of taxpayers and help tax reviewers solve problems encountered during tax audits to prevent tax evasion. The method combined regression with an SVM and prioritized high-income taxpayers when selecting data.

Zhang [29] proposed a detection model for enterprises' false issuance of VAT invoices based on an SVM. The researcher set a risk coefficient and divided enterprises into different levels to indicate their likelihood of false invoice issuance.

The idea behind an SVM is simple. Its classification effect on small-scale data is good, and its generalization ability is strong. However, it also has high computational complexity and is sensitive to missing data when dealing with large-scale data. Moreover, SVMs generally struggle to perform well on datasets where the imbalanced ratio is very large [30].

3.1.4. Bayesian classifier

A Bayesian classifier is a statistical inference process based on Bayes' theorem [31,32]; it uses prior information about the parameters and the likelihood probability calculated by the statistical model of the existing data to obtain the posterior probability of an unknown parameter. Bayesian classifiers are an essential technique in statistics, especially for the dynamic analysis of sequence data, and are widely used in science, engineering, medicine, and other fields.

Kirkos et al. [33] compared three data mining techniques—namely a decision tree, neural network, and Bayesian belief network—to detect fraud in financial statement data. Through experimental verification, the Bayesian belief network model was determined to have the best performance, achieving an accuracy of 90.3%. This study can help authorities detect financial statement fraud.

Kang et al. [34] constructed a tax assessment method based on a Bayesian classification model and conducted an empirical analysis with real data. The results showed that the Bayesian classification tax assessment model can be broadly applied. Zhang et al. [35] proposed a method to detect tax declaration fraud based on a Bayesian classifier, which determined whether the tax declaration amount of an enterprise was abnormal. Lenz et al. [36] also studied the application of the Bayesian method to detect tax fraud and verified the method on cases in Germany.

Bayesian classifiers have a relatively solid theoretical foundation [37]; moreover, they are stable and less sensitive to missing data. The disadvantage is that the sample attribute independence assumption is employed; thus, the effect will be limited when the sample attributes do not satisfy the independence assumption. It should be noted that there is often a certain correlation between the selected attributes in tax scenarios.

3.1.5. Logistic regression

LR is a form of multivariate analysis [38]. It is a commonly utilized method in computer science, econometrics, biostatistics, and other disciplines [39]. In brief, LR is a machine learning method for solving binary classification problems and is used to estimate the likelihood of something occurring.

Qi et al. [40] established an index system based on actual cases. The researchers used an LR model to discriminate between tax cases, which improved the accuracy and efficiency. Wang et al. [41] selected nine financial indicators for the selection of tax cases, regarded tax auditing as a two-category problem, and employed LR models to predict the types of cases. Su [42] took some wholesale enterprises in Guangzhou as samples and established an LR model for the tax review of the wholesale industry. The results of their study showed that, in order to reasonably explain tax compliance behavior, it is necessary to use indicators with significant characteristics. Using this approach, different tax review models can be constructed for different regions and industries to achieve the best results. Yuan [43] summed up five tax evasion methods commonly used by taxpayers and used an LR model to construct a tax review model for identifying corporate tax evasion in a given city. The accuracy of this model reached 79.5%.

LR is a simple and efficient algorithm with low time and space complexity. In essence, however, LR is also a linear classifier;

accordingly, it is difficult to find correlations between features in a high-dimensional feature space, which will be prone to underfitting [44].

3.1.6. Clustering models

Most existing work in the use of artificial intelligence for tax risk detection is based on supervised or semi-supervised machine learning techniques. However, auditing tax-paying companies is a slow and costly process. Therefore, there is a limited amount of available historical tax evasion information, especially labeled data, which severely hinders the use of certain supervised machine learning models in tax risk detection. This has motivated some researchers to employ clustering models [45,46], an unsupervised method, to identify tax risks.

Denny et al. [47] proposed an SOM-based visualization method to mine abnormal hotspot clusters in a customer dataset of the Australian Taxation Bureau. An SOM maps topological maps from high-dimensional data to a 2D map, in which similar entities tend to be closer together. SOMs also provides a variety of visualizations to enable non-technical users to explore the dataset and to help analysts with policy formulation.

Liu et al. [48] used hierarchical clustering in the selection of tax audit cases. Hierarchical clustering analysis was carried out on the index data of 30 enterprises. The obtained analysis results were then compared with known tax evasion cases, so as to assist in case selection and improve the efficiency and effect of tax auditing. Liu et al. [49] proposed a clustering-based data mining algorithm to mine the outlier problem in the tax industry. The proposed approach can find abnormalities in massive tax data; moreover, it can not only filter key data sources but also be used to detect taxpayers' abnormal business behavior. It can even determine whether taxpayers are suspected of tax evasion and tax fraud, enabling it to quickly and accurately identify candidates for tax audits.

Assylbekov et al. [50] proposed an unsupervised method based on a Kohonen SOM to detect VAT evasion behavior among legal entities in Kazakhstan. The results showed that this method is superior to the current scoring model used by the National Tax Commission of the Republic of Kazakhstan. De Roux et al. [51] proposed a novel unsupervised clustering-based method to detect potential fraudulent behavior among taxpayers. The experimental results on 1367 tax returns in Colombia showed that the operational efficiency of the tax supervision process can be improved without the need for historically labeled data. Xia et al. [52] proposed an improved *K*-means clustering algorithm. This approach, which involved an unsupervised learning model, was used to analyze and evaluate tax risk in the equity transfer of real estate enterprises. The experimental results verified the effectiveness of the method.

The clustering algorithm is simple and easy to implement; moreover, it does not require label information, making it highly suitable for tax scenarios. Thus, it has attracted extensive research attention. However, clustering algorithms are often sensitive to noise and outliers [53,54], and they converge very slowly on large-scale datasets.

3.2. Non-relationship-based deep models

3.2.1. Artificial neural networks

Neural networks, which belong to the field of machine learning and cognitive science, are mathematical models that imitate the structure and function of biological neural networks [55]. A typical neural network has three elements: structure, activation function, and learning rules. In recent years, as computing power and the completeness of theoretical foundations have improved, the depth limitation of neural networks has been alleviated, and newly proposed deep learning methods have rekindled the focus of academia and industry on neural networks. At present, neural networks are widely used in computer vision [56], data mining [57], machine translation [58], and other fields [59].

Li et al. [60] combined fuzzy logic theory and neural networks, giving full play to the advantages of the two algorithms, which include the ability of fuzzy logic theory to adjust to specific problems and the nonlinear approximation ability of neural networks. The researchers first constructed a tax credit evaluation index system, and then established a tax evaluation model based on a fuzzy neural network. Finally, they obtained results through simulation and analysis, which showed that the model performed well on the task of credit evaluation for taxpaying enterprises.

Lin et al. [61] combined expert questionnaires and data mining techniques to rank the importance of fraudulent elements in financial statements, and then classified and identified statement fraud. The data mining techniques used in their study included an artificial neural network, LR model, and decision tree. The experiments showed that the effect of the artificial neural network was better than that of the decision tree or LR model, with the former obtaining a classification accuracy of 92.8%. The researchers contended that neural network-based discrimination methods can efficiently assist auditors in completing their work.

Assylbekov et al. [62] proposed a method based on mathematical statistics to detect VAT evasion. In their study, an SOM was used to conduct a preliminary analysis of the data, so as to judge the nature of taxpaying enterprises. SOMs are a type of artificial neural network that use unsupervised methods to process data and obtain subsequent representations. This approach not only preserves the topological properties of the input space but is also applicable to the visualization of high-dimensional data in a low-dimensional space. The experimental results showed that the proposed model outperformed its competitors.

Lopez et al. [63] used a neural network to calculate the probability of tax evasion among taxpayers. The researchers validated the model on Spanish tax data and found that its accuracy reached 84.3%. Zhang et al. [64] proposed a machine learning-based tool to help tax authorities detect transaction-based tax evasion on social media platforms. More specifically, a multi-modal deep neural network was used to automatically detect suspicious behaviors on the platforms, enabling the identification of e-commerce-based tax evasion on a large scale. Chen et al. [65] proposed a tax evaluation method based on an artificial neural network, while considering the fact that the characteristics of the neural network enable it to adapt effectively to nonlinear tax data.

Zhang et al. [66] developed a Regtech tool that automatically detects transaction-based tax evasion activities on social e-commerce. They collected a dataset and manually annotated sampled posts with multiple labels related to sales and tax evasion activities. Then, they developed a multi-modal deep neural network model to automatically detect transaction-based tax evasion activities from the posts. The experimental results showed that the performance was superior to those of any single modality models.

Murorunkwere et al. [67] used an artificial neural network to detect income tax fraud in Rwanda. The model achieved high accuracy, precision, and recall score, and the evidence from the study can help auditors reduce audit time and cost and recover lost revenue. Mojahedi et al. [68] used an improved particle swarm optimization (IPSO) algorithm to optimize multilayer perceptron (MLP) neural network and SVM classifiers. IPSO-MLP and IPSO-SVM models using the IPSO algorithm were used as new models for tax evasion detection. The proposed system was evaluated on a dataset collected from the general administration of tax affairs in the West Azerbaijan province of Iran, with 1500 samples. The experiments showed that the IPSO-MLP model outperformed other models, with an accuracy rate of 93.68%.

Alsadhan et al. [69] presented an anomaly-detection technique for identifying tax fraud without requiring historically labeled data. The technique uses stacked autoencoders (SAEs) to compare the probability distributions of suspicious values for each field on the tax return form. The results showed that the proposed method is effective in identifying current tax fraud schemes. Potential limitations and future extensions were discussed, such as adding additional financial ratios and growth-related features, and incorporating the opinion of tax experts when calculating the anomaly score. The researchers suggested a supervised auditing strategy combining supervised and unsupervised detection methods for an optimal auditing strategy.

Neural networks have achieved superior performance in many fields, and their powerful performance is an important reason for the rise of the third wave of artificial intelligence. However, some problems remain with artificial neural networks, including the need for a large number of labeled training samples to train the model and poor model interpretability [70,71]. Unfortunately, data labels are particularly difficult to obtain in tax scenarios [72], as labeling in this context requires a great deal of expert knowledge. In addition, model interpretability is especially important in the tax field, as it can provide clues for auditors.

3.2.2. Hybrid models

Some studies use ensemble learning, or divide tasks into multiple stages, to integrate models for risk detection. A hybrid model is a method that combines two or more different models to take advantage of their respective strengths and compensate for their weaknesses in order to achieve good recognition results [73].

Ravisankar et al. [74] used various data mining techniques (e.g., multilayer feed forward neural network (MLFF), an SVM, genetic programming (GP), a group method of data handling (GMDH), LR, and a probabilistic neural network (PNN)) to analyze financial statement fraud among 202 Chinese companies with 35 financial items. The results showed that PNN and GP achieved particularly good performance.

Zheng et al. [75] applied decision tree and regression modeling methods to construct a regression model for taxpayers' evasion risk rating, using cluster analysis, outlier analysis, and association rules to build an industry transaction rule base. Gonzlez et al. [76] used clustering algorithms such as an SOM to identify groups of taxpayers with similar behaviors. Subsequently, they used decision trees, neural networks, and Bayesian networks to identify variables associated with fraudulent or non-fraudulent behavior and detect patterns in related behaviors, thereby generating knowledge that could help Chilean tax authorities identify fraud and accordingly detect tax crimes.

Song et al. [77] proposed a hybrid method for assessing financial statement fraud risk by combining machine learning methods with rule-based systems. The machine learning model adopts an integrated learning model comprising four models: LR, a neural network, a decision tree, and an SVM. When applied to the data of 550 companies between 2008 and 2012, the method outperformed machine learning methods in assessing the risk of financial statement fraud. Rahimikia et al. [78] studied a hybrid intelligent system that combines MLP, SVM, and LR classification models with a harmony search (HS) optimization algorithm to detect the effectiveness of corporate tax evasion in data obtained by Iran's State Taxation Administration (INTA). In the food and textile industries, the accuracy rate reached 90.07% and 82.45%, respectively. The tax evasion detection method based on positive and unlabeled learning (TEDM-PU) proposed by Wu et al. [79] utilizes limited labeled data and a large amount of unlabeled data to detect tax evasion. This method applies a random forest model to preprocess tax data, then assigns pseudo-labels to unlabeled samples based on positive and unlabeled (PU) learning; finally, it uses LightGBM for tax evasion detection.

Javadian et al. [80] proposed an improved iterative dichotomizer 3 (ID3) decision tree model combined with a multi-layer perceptron neural network that was optimized via a genetic algorithm to improve its performance and accuracy. Their study on the financial statements of companies listed on the Tehran Stock Exchange verified the validity of the proposed model. Rahman et al. [81] tested five machine learning algorithms (LR, nearest neighbor algorithm, naive Bayes, decision tree, and random forest) on a real dataset of 3365 companies listed on the Malaysian stock exchange from 2005 to 2015. Mekonnen et al. [82] used data mining to detect and predict tax evasion by taxpayers in Addis Ababa, Ethiopia. Following the Cios model, the researchers developed a cluster model using a K-means algorithm and a classification model using an MLP algorithm with PART rule. The model achieved high accuracy and identified important variables such as tax, liability, and expenses. Savić et al. [83] proposed the HUNOD method, which combines K-means and an autoencoder for robust outlier detection and uses a decision tree to obtain an explainable surrogate model for detected outliers. The method was evaluated on two datasets from the Tax Administration of Serbia, and the results showed that it could identify 90%–98% of internally validated outliers, depending on the clustering configuration and regularization mechanisms used.

Baghdasaryan et al. [84] explored the use of machine learning tools—more specifically, gradient boosting—to develop a fraud prediction model for Armenian business tax payers. A gradient boosting machine builds models sequentially, with each subsequent model (i.e., decision tree) aiming to reduce the error of the previous ones. Features are taken in different subsets by each node to find different signals from the data. The model successfully derived important features from tax returns, including historical fraud and auditing, share of administrative costs, and external economic activity. The researchers demonstrated that even moderately accurate models can improve the existing accuracy of rule-based approaches and that information contained in the supplier and buyer network of the taxpayer can be used as predictors of fraud, which is particularly useful for newly established companies.

As hybrid models [85] are composed of multiple models, they can combine the advantages of each model to achieve better recognition results while enhancing the robustness and reliability of the model. However, there are certain drawbacks associated with hybrid models, such as a need for more computational resources and a more complex parameter tuning process. In addition, hybrid models require more data for training and testing to achieve better performance, rendering them less than ideal for low-resource situations, such as taxation, where acquiring data labels proves to be more challenging.

It is worth noting that the labeling of large amounts of data by tax experts is a long and expensive process. Moreover, differences in economic policies and patterns across regions lead to different characteristics of data distribution. Existing models often do not consider cross-regional use. As a result, some hybrid models based on the idea of transfer learning have emerged.

Zhu et al. [86] proposed the new inter-regional tax evasion detection method based on transfer learning (IRTED-TL). By combining feature-based and instance-based transfer learning, IRTED-TL can obtain supplementary knowledge of source regions and sufficient training data, which is then applied to target regions with sparse labels in order to augment training data in the presence of regional differences. Wei et al. [87] proposed the general unsupervised conditional adversarial network (UCAN) architecture and applied it to cross-regional tax evasion detection. This architecture uses the labeled data of other audit tasks to assist in target audit tasks with sparse labeling and reduce intra-class distribution

differences. End-to-end learning for unsupervised feature transfer is achieved by combining a distribution adapter with a label predictor. Zhang et al. [88] proposed the transferable tax evasion detection method based on positive and unlabeled learning (TTED-PU). This method combines PU learning with deep transfer learning to solve the transfer problem of marginal probability distribution and conditional probability distribution in the transfer process. Transfer learning is highly suitable for low-resource scenarios such as taxation and can utilize data effectively [89]. However, there are still some problems with this approach: First, it is difficult for the algorithm to converge, as the source domain and target domain may be different; second, the algorithm in the target domain may inherit the defects of the source domain [90].

3.3. Non-relationship-based reinforcement learning models

3.3.1. Value-based reinforcement learning

Reinforcement learning [91–93] emphasizes interaction with the environment through learning strategies to maximize the expected benefits. However, it is difficult to design the reward function and ensure that the reinforcement learning algorithm converges. Relatively few studies have been conducted on risk detection based on reinforcement learning.

Abe et al. [94] proposed a new reinforcement learning framework based on a constrained Markov decision process, which closely combines data modeling and optimization techniques. The researchers deployed this system at the New York State Department of Taxation and Finance. Goumagias et al. [95] combined deep reinforcement learning and Q-learning to determine expected tax evasion risk behavior among taxpayers. The researchers took the Greek tax system as a specific case study and reported on the relevance of issues regarding the expected behavior of companies, incentives for profit reporting, risk aversion, and policy implications.

To sum up, reinforcement learning [96,97] has the advantage of an optimized performance and the ability to sustain change for a long period of time. However, it also carries the risk of state overload, which can have a negative impact on outcomes. Moreover, it requires substantial amounts of data and computational resources to perform effectively. It can only showcase its unique strengths in solving complexity that is intractable through other means when possessing enough computing power and data, which is not always possible in tax scenarios.

3.4. Other non-relationship-based models

3.4.1. Evolutionary algorithms

Evolutionary algorithms [98,99] are a branch of the evolutionary computing field. The mechanism of an evolutionary algorithm is inspired by biological evolution and is designed to find the optimal solution in the solution space by simulating the process of biological evolution. Alden et al. [100] used a genetic algorithm and an estimation of distribution algorithm to train fuzzy rule-based classifiers for financial fraud pattern detection. Their results showed that the genetic algorithm and estimation of distribution algorithm could classify the unseen data of enterprises more efficiently than traditional LR models and could effectively detect financial fraud. The classification accuracy of the two algorithms reached 75.47% and 74.26%, respectively, through 10-fold cross-validation. Warner et al. [101] proposed a prototype evolutionary algorithm that takes asset types, tax entities, and rule sets for transactions between entities as inputs. The algorithm provides auditors with scheme information on tax evasion. These schemes are ranked using a "fitness function," and the best scheme receives the highest tax deduction and the lowest penalty.

Hemberg et al. [102] suggested that tax evasion schemes and audit procedures are competitive relationships. This mutual influence aligns well with the nature of evolution. The researchers proposed a co-evolutionary model that models the transaction sequences of taxpayer networks and censorship. The model helps tax agencies model how changes in tax laws or censorship might drive change in tax evasion schemes. In addition, in the following year, Hemberg's team [103] proposed the simulating tax evasion and law through heuristics (STEALTH), which was designed to predict tax evasion by modeling the co-evolution of tax evasion schemes with censorship. The researchers explored the tax schemes that emerged in response to changes in audit procedures. Finally, experiments verified the feasibility of using the method to detect tax evasion.

Evolutionary algorithms are inspired by biological evolution, which is readily comprehensible. However, these approaches also have many parameters, and parameter selection is often reliant on experience. Inappropriate parameters lead to slow convergence and have a serious negative impact on the results [104–106].

3.4.2. Agent-based models

Agent-based models [107–109], also known as multi-agent systems, are computational models used to simulate the actions and interactions of independent individuals with self-consciousness. Their purpose is to evaluate the role of individuals in the system. Agent-based models are usually used in computer science, economics, social sciences, biology, and other fields. Such models can simulate complex phenomena, although they require many parameters to be preset.

Antunes [110] used an agent-based model to explore the reasons for tax evasion. The researchers argued that the agent-based model is effective in exploring tax compliance issues because it can provide empirical explanations for decisions based on individual motivations. In addition, the agent-based model can explore individual psychology, agent interaction, and social mechanisms. This strong explanatory power can be used to predict the future of the social system, which can in turn be used to design a tax regulation system that reduces tax evasion. The study experimentally demonstrated that a personalized agent-based model can help auditors to examine tax evasion issues more effectively.

Lima et al. [111] used an agent-based Monte Carlo simulation method and added a noisy majority voting method to the Zaklan model, which was then applied to tax evasion detection on the Apollo network. The purpose of this study was to test the robustness of the Zaklan model on the Apollo network, as well as to verify the results of previous tax evasion detection based on this method. The experiments showed that the majority voting-based Zaklan model proposed in this study is highly robust and can provide effective assistance for tax evasion auditing. In addition, the researchers contended that the higher the probability of a review hit is, and the stronger the penalty is, the less tax evasion will occur.

Llacer [112] proposed a novel agent-based model, SIMULFIS, which simulates tax rules and tax evasion. The research explores three distinct aspects: Theoretically, it studies the interrelationship between the factors underpinning tax evasion (utility maximization, fairness, social impact); methodically, the model constructed by the agent-based model is relatively realistic (e.g., by ascribing unique characteristics to individuals); and politically, models that are proven to be valid are more likely to be useful tools for assessing existing tax policies and detecting tax evasion. The model was applied to the actual case of Spain. The results were basically consistent with the theoretical expectations, meaning that the reliability of the model is supported.

Noguera et al. [113] contended that, in the field of tax behavior analysis, agent-based models are highly effective at testing

theories and hypotheses. Their study accordingly proposed an agent-based simulation model of taxation rules, which combined rational choice and the rules of social influence to generate an aggregation model of taxation behavior. The potential of the model was then experimentally verified.

Andrei et al. [114] argued that agent-based models are not only flexible but also have strong analytical capabilities and achieve good results when applied to complex issues such as tax rules and tax evasion mechanisms. Their research showed that a network structure has a significant impact on the dynamics of tax rules, demonstrating that taxpayers closer to the center of the network are more willing to declare all of their income, especially when faced with large penalties. This work also revealed that a network structure should be considered an important factor when modeling tax rules, as different topologies may lead to different results. Notably, the model proposed in this study is more stylized and abstract than tax evasion in the real world. The researchers plan to optimize the model's performance by incorporating the socioeconomic and political context of specific cases.

Bloomquist et al. [115] compared three multi-agent-based personal income tax evasion detection models. The researchers discussed the similarities and differences between these models and the advantages of multi-agent-based models in the field. The similarities lie in the homogeneity of these models, the number of agents instantiated, and so forth, while the differences lie in the external data used to validate the models, the ability to estimate the effects of audits, the characteristics of the agents, the specific implementation, and so forth. Following a comparative analysis, the researchers concluded that the most important inspiration from their study was that process validity is very important when developing computational models for policy analysis. They further highlighted the importance of avoiding the use of black-box techniques when conducting research; despite such techniques being popular, they can confuse experts examining situations from a practical perspective.

Agent-based models [116,117] possess several advantages, such as the ability to model heterogeneous populations, the ability to model complex systems with minimal coding, and the ability to simulate the actions and interactions of independent individuals with self-awareness. However, there are also some drawbacks to agent-based models, such as the high computational requirements, their sensitivity to parameter values, and the difficulty in validating results. Unfortunately, such models require many parameters to be preset when simulating complex phenomena such as taxation. The dynamics of agent-based models depend on determined parameter values of the agents' rules and attributes, and different settings may lead to wholly distinct behavioral patterns and system properties. Finding a reasonable and stable parameterization can be a formidable task. In addition, as they are rule-based simulations, it is challenging to determine whether they produce emergent behaviors and properties that match the real world.

Table 3
Summary of non-relationship-based risk detection methods.

Method	Advantages	Weakness	References
Association rule	Simple; results are easy to understand	As the amount of data increases, the amount of computation increases rapidly	[8,9]
Tree-based	Easy to understand and interpret; ability to handle missing attribute samples	Prone to overfitting; difficult to support online learning	[13–20]
Support vector machine	High classification accuracy and generalization ability in small-sample cases	Large-scale training data is computationally intensive and sensitive to missing data	[24–29]
Bayesian classifier	Strong mathematical theoretical foundation and robustness	Reduced performance when sample attributes do not satisfy independent and identically distributed assumptions	[33–36]
Logistic regression	Simple and efficient, with low computational complexity and a low storage footprint	Prone to underfitting	[40–43]
Clustering model	No label information is required; the algorithm is easy to implement	Slow convergence of large datasets; sensitive to noise and isolated points.	[47–52]
Artificial neural network	High accuracy and easy parallel processing	Large training sample requirement and limited interpretability	[60–69]
Hybrid model	Strong robustness and leverages the benefits of multiple models	Difficult training process; difficulties with convergence; requires more data.	[74–84,86–88]
Reinforcement learning	Modeling of sequential decision problems and consideration of long-term rewards	Reward function design is difficult; convergence is unstable; risks state overload; requires more data and computational resources	[94,95]
Evolutionary algorithm	Robustness and parallelism	Too many control variables; slow convergence	[100–103]
Agent-based	Simulation of complex phenomena is possible	Many parameters need to be preset; high computational requirements; difficulty in validating results	[110–115]

3.5. Summary

Existing non-relational-based risk detection methods are summarized above. The advantages and disadvantages of each technique and the specific literature are analyzed and listed in Table 3 [13–20,24–29,33–36,40–43,47–52,60–69,74–84,86–88,94,95,100–103,110–115].

4. Relationship-based tax risk detection methods

The aforementioned non-relationship-based risk detection approach starts from the individual taxpayer and often requires the manual selection, design, and construction of different features based on the experience of tax experts. A non-relational risk detection model is then trained using these features, along with different training paradigms, and is finally applied to tax risks. It is notable that the non-relationship-based risk detection approach can lead to the loss of a large amount of interaction information. However, the tax scenario can be naturally constructed as a complex network that contains a range of entity objects, such as tax enterprises, legal persons, commodities, and tax laws and regulations. There are also various types of relationships between the different entities in play, such as transaction relationships between tax enterprises, investment holding relationships between tax enterprises and legal persons, enterprises buying and selling commodities, and so forth. Tax evasion behaviors often exhibit

"gang-like" characteristics, such as tax evasion through related transactions. Identifying such risky behaviors requires more consideration of the importance of relationships and the mining of tax evasion clues from networks with richer semantic information.

4.1. Graph pattern matching

Graph pattern matching [118,119] is one of the most important research avenues related to graphs; thus, it has attracted extensive research attention in fields such as data mining and databases. In tax scenarios, researchers typically use graph pattern-matching algorithms to mine tax evasion groups across the entire tax transaction network.

Tian et al. [120] proposed a colored network-based model (CNBM) to characterize the economic behavior, social relations, and interest-related transactions among taxpayers, and accordingly generate a taxpayer interest interaction network. Suspicious groups in interest interaction networks are discovered by building pattern trees and matching component patterns. Wei et al. [121] proposed a new graph-based suspicious groups of interlock-based tax evasion detection method, named GSG2I, which includes a graph projection algorithm designed to identify relationships that satisfy controller interlock patterns and component-based pattern matching. The algorithm finds suspected tax evasion groups based on controller interlocks. Experimental tests on seven years of tax data in a Chinese province showed that the GSG2I method can greatly improve detection efficiency.

Liu et al. [122] designed a false VAT invoicing behavior detection system and proposed a depth-first-based directed graph loop search algorithm that detects fund loops in fund transaction flow graphs and can query the details of the loop-associated accounts to save audit costs. Ruan et al. [123] proposed a hybrid method based on tax rate difference detection, topological pattern matching, and tax anomaly detection to identify affiliated-transaction-based tax evasion.

Tax evaders perform so-called circular trading by invoicing sales between groups that add no value and are not associated with any actual supply of goods, contributing to the commission of a variety of financial crimes. Mathews et al. [124] developed a circular trade model for the commercial taxes department of the government of Telangana, India, involving three dealers. The model could predict whether future links would form between two dealers leading to the creation of a triple loop with an accuracy of 80%.

Rocha et al. [125] presented an innovative methodology for the detection of shell companies in financial systems using legal person attributes and dynamic social networks. The proposed model outperformed the traditional rules method in terms of balanced accuracy, true positive rate, and false-positive rate. The technique has been successfully implemented in a Mexican financial company and could be applied by other financial institutions to identify shell companies and reduce the prevalence of tax avoidance and money laundering. The limitations of this study included not regularly owning with confirmed cases of shell companies and not considering connections with clients in other financial institutions. Future research should focus on modeling suspicious internal connections and connections between an internal legal person and an external one.

Chen et al. [126] proposed a novel framework called AntiBenford subgraphs for unsupervised anomaly detection in financial networks. The framework is based on statistical principles and can efficiently find anomalous subgraphs in near-linear time. The framework was evaluated on real and synthetic data and exhibited superior performance compared with state-of-the-art graph-based anomaly detection methods. The proposed AntiBenford subgraphs exhibit the characteristics of illicit transactions and can provide novel insights into financial transaction data. The paper concluded by suggesting future directions for research, including the design of algorithms for overlapping anomalous subgraphs and the inclusion of other measures of statistical deviation in the experimental setup.

The results of graph pattern matching are easy to understand, and visual analysis methods are often employed to further analyze the pattern-matching results. However, the subgraph matching problem is too computationally intensive when the scale of data is large [127–129]. Moreover, matching graph patterns often need to be manually defined; in addition, there is subjectivity and lag in the tax audit game process, which can result in many important graph patterns being overlooked.

4.2. Graph representation learning

Graph pattern matching approaches are heavily reliant on the experience of tax experts to summarize and extract the tax evasion patterns; when new patterns appear, they need to be recorded via hard coding. Meanwhile, the basic characteristics of tax-paying enterprises are not fully considered. In order to solve this problem, researchers have begun to study the tax evasion detection problem by using graph representation learning [130,131].

Matos et al. [132] introduced a novel feature selection algorithm based on complex network techniques, which can capture key fraud indicators. A classifier for accurate tax fraud detection using the above algorithm has also been proposed. The effectiveness of the algorithm was verified on a real dataset obtained by the State Treasury Office of Brazil.

Wu et al. [133] proposed a new tax evasion detection framework based on fused transaction network representation (TED-TNR). This approach jointly embeds the topological information of the transaction network and the basic attributes of taxpayers into a low-dimensional vector space, and then exploits the low-dimensional vectors of taxpayers for tax evasion detection. The results showed that the TED-TNR method can detect tax evaders more accurately than existing methods.

Mi et al. [134] proposed a tax evasion detection method based on PU learning with Network Embedding features (PUNE). First, the transaction network features are extracted by means of the network embedding technique. Second, individual weights are assigned to each sample based on class priors and sorted rank scores in the PU learning process. Finally, a weighted sample classifier is trained based on minimizing the empirical risk.

An et al. [135] proposed a method that relies on network embedding based on upstream and downstream for tax risk identification (NEUD-TRI). The method designs optimization functions to capture local and global static and dynamic network structures, respectively. Empirical results on a provincial tax dataset confirmed the validity of the model.

Wang et al. [136] proposed the temporal edge-enhanced graph attention network (T-EGAT) method. In this method, edge-enhanced graph attention networks are used to learn complex topologies and thereby capture spatial dependencies, while recurrent weighted average units are used to learn the dynamics of transaction data in order to capture temporal dependencies. Experimental tests on tax data showed that the method outperformed existing methods when detecting tax evaders.

Gao et al. [137] proposed a multi-stage tax evasion detection framework named FBNE-PU. This framework significantly improves the tax evasion detection performance by extracting effective features from the transaction network and leveraging a small number of positive samples along with a large number of unlabeled samples. Shi et al. [138] proposed a novel graph neural network model, named Eagle, for tax evasion detection using a heterogeneous graph model. Based on the guidance of designed metapaths, Eagle can extract more comprehensive features

through a hierarchical attention mechanism that fully aggregates taxpayers' features with their relations. Extensive experiments on a real-world tax dataset showed that Eagle outperforms state-of-the-art tax evasion detection methods in both classification and anomaly detection scenarios.

Graph representation learning makes use of large amounts of structural and relational information, and often achieves better risk detection performance. However, when the graph scale increases, problems arise related to poor interpretability and increased computational complexity, which urgently need to be solved [139].

4.3. Visual analysis

Visual analysis, which is based on an interactive visual interface for visual presentation, helps researchers to understand and further implement methods of analysis and reasoning [140–142]. In the field of tax risk auditing, most existing data mining methods lack interpretability, making it difficult to provide direct evidence of tax evasion. It is hard for tax inspectors to understand and trace the source of a finding based solely on the algorithm's classification results. The visual analysis method constructs a complex network of entities, such as taxpayers and their transactions and interests. Because the results are easy to understand, visual analysis has become an indispensable technology in tax auditing.

Didimo et al. [143] designed a visual analysis system of financial activity networks named VISFAN, which combines social network analysis and cluster analysis for financial transaction networks, with the goal of detecting financial crimes such as money laundering and fraud. Tselykh et al. [144] used clustering and rule induction techniques to identify potentially fraudulent transfer pricing risk behaviors in attribute graphs. Their approach used network analysis and visualization methods to screen out entities that required special attention for transfer pricing audits.

The Italian Revenue Agency [145] developed TaxNet, a decision support system for tax evasion detection based on visual analytics. This system allows users to intuitively define and extract suspicious patterns in taxpayer networks. The system is currently in use at the tax office in the region of Tuscany and has demonstrated its effectiveness in a real working environment. Zheng et al. [146] designed ATTENet, a visual analysis system for detecting and interpreting suspicious affiliated transaction-based tax evasion (ATTE) groups. The suspected value of tax evasion groups can be detected by the network embedding method Structure2Vec [147] and a random forest algorithm, after which the detection results are visually explained.

Yu et al. [148] designed TaxVis, a visual inspection system for tax auditors. The system performs tax evasion group detection according to a two-stage approach. In the first stage, the network embedding method node2vec [149] is used to learn the representation of embedding enterprises from the corporation-associated network, and the suspicious score of each corporation is calculated using LightGBM. In the second stage, the system uses visualization methods such as Sankey diagrams to analyze the abnormal upstream and downstream transactions of suspicious companies.

Didimo et al. [150] proposed a new approach called MALDIVE (MAtch, Learn, DIffuse, and VisualizE) to detect tax risk behavior among taxpayers through graph pattern matching, social network analysis, and machine learning. An information diffusion strategy is also used to expand the set of possible risky taxpayers. The results are finally output to tax inspectors using a network visualization system.

Zha [151] used hierarchical convolutional networks to calculate the risk scores of taxpayers in a constructed tax audit network. A visual analysis system, TaxAA, was designed to allow tax auditors to customize suspicious indicators and ultimately observe suspicious relationships among taxpayers in the form of a "wheel" diagram. Lin et al. [152] proposed an interactive visual analysis system named TaxThemis that helps tax officials to mine and explore suspected tax evasion groups through the analysis of heterogeneous tax data. A new coding scheme was proposed to visualize evidence of income transfers through related party transactions in a calendar heatmap.

Visual analysis is very easy to understand due to its visual presentation and has the advantage of high interpretability. However, visual information is often designed by data scientists based on their own experience; as a result, the data presented tends to be subjective and biased [153–155], which may result in important clues being overlooked and incorrect risk detection judgments being made.

4.4. Summary

In the previous sections, we discussed existing relationship-based tax risk detection techniques. Each tax risk detection method has its own unique advantages and disadvantages. We analyze the advantages and disadvantages of each technique and list the relevant literature in Table 4 [120–126,132–138,143–146,148,150–152]. These findings will aid tax data scientists in selecting appropriate tax risk detection techniques.

5. Open issues and future research directions

Taxation is the foundation of a nation. Due to the importance of taxation, tax risk detection has long been an important research topic. In recent years, great progress has been made in relevant research works both at home and abroad, and numerous excellent tax risk detection methods have emerged. However, the existing methods remain data-driven and impacted by certain limitations, including the fragmentation of knowledge (making it complex to integrate and utilize), tax risk detection results being difficult to explain, tax risk detection algorithms being computationally expensive, and algorithms being reliant on label information manually provided by tax experts. It is difficult to solve the above theoretical and technical problems by relying on data-driven methods alone. The use of knowledge-guided and data-driven big data knowledge engineering [156–158] will be the future development trend in the field of tax risk detection, as the only path from informatization to intelligence. In this section, we specifically discuss the four limitations in existing works mentioned above and look ahead to future tax risk detection methods based on big data knowledge engineering.

Table 4
Summary of relationship-based risk detection methods.

Method	Advantage	Weakness	References
Graph pattern matching	Results are easy to understand	Computationally intensive; relies on manually designed tax evasion group patterns	[120–126]
Graph representation learning	High accuracy and strong generalization ability	Poorly interpretable and computationally intensive when graph scale increases	[132–138]
Visual analysis	Easy to understand and highly interpretable	Subjective and biased	[143–146,148,150–152]

5.1. Research direction 1: Fragmented knowledge fusion based on big data knowledge engineering

Existing tax risk detection methods often start from the registration and invoice information provided by taxpayers to construct features for risk detection. However, knowledge in real-world tax environments is multi-source, multi-domain, and multi-modal; moreover, some data are still not fully utilized, such as documents issued by the state administration of taxation and local taxation departments, relevant laws, regulations, and policies, and third-party information (e.g., existing case descriptions and public security information pertaining to taxpayers). Therefore, exploring how to transform multi-source heterogeneous and fragmented data into a machine-representable and computable structured knowledge base through fragmented knowledge fusion in tax scenarios is challenging yet necessary work that must be pursued in the future. Fragmented knowledge fusion also satisfies the multiple knowledge representation (MKR) framework [159], which enhances the robustness and interpretability of a model by aggregating information from multiple sources, thus achieving more intelligent applications, such as taxable tax calculations [160].

Focusing on the above problems, it will be necessary to carry out the work from the following two aspects: ① For semi-structured and unstructured data in the fiscal and taxation fields, such as tax policies and regulations, there is a need to study knowledge extraction, entity extraction, relationship extraction, and attribute event extraction, in order to transform these data into structured knowledge; and ② in the context of knowledge fusion, there is a need to study co-reference disambiguation and entity linking. At the same time, eliminating the problems of large differences between domains and variable data distribution in multi-source heterogeneous data will be the key to realizing the fusion of fragmented knowledge.

5.2. Research direction 2: Interpretable cognitive reasoning based on big data knowledge engineering

Most existing tax risk detection methods are so-called "black-box" models with poor interpretability; such models can only know the "hows" but not the "whys" and thus cannot directly provide relevant evidence for use in uncovering corporate tax evasion. Visual analysis-based methods can mine and explore individuals or groups suspected of tax evasion through machine learning algorithms, and employ an easy-to-understand visual interface to help tax officials select the appropriate cases. Although this approach significantly improves interpretability, it still uses a black-box model. In addition, the data displayed by visual analysis is likely to be biased and subjective, impacted by the interpretations of data scientists, which will lead to important clues being ignored. Future research in this area must focus on how to carry out high-level cognitive reasoning in the tax knowledge base obtained by using fragmented knowledge fusion, with the use of cognitive reasoning to generate an interpretable and complete evidence chain, and to help tax inspectors trace the sources of suspicion.

To address the above problems, the following research directions are recommended: ① Exploring the interpretable cognitive reasoning of transformer-based methods in tax risk identification, as transformer-based models have stronger expressive power, while self-attention-based mechanisms have better interpretability for presenting risk relationships among entities, making this direction extremely attractive and meaningful; ② Using different paradigms (e.g., transductive learning, inductive learning, and deductive inference) to expand and evolve existing knowledge; ③ Combining the complementary information of symbolism and connectionism, and using existing knowledge to guide data reasoning, so as to generate an evidence chain related to tax evasion and fraud and thereby assist tax inspectors in tracing the source.

5.3. Research direction 3: Risk detection methods for large-scale tax scenarios

Existing tax risk detection technologies tend to focus on the accuracy of risk detection and often improve this accuracy by employing ensemble learning and building larger and more complex models. However, China is home to hundreds of millions of taxpayers who issue tens of billions of invoices every year. In addition, the tax scene must be combined with third-party data, such as industrial and commercial data, customs and public security data, and so on. In real-world tax scenarios, it is necessary to deal with extremely large-scale data, which may lead to the failure of many complex risk detection models that cannot be used directly. Designing a universal tax evasion and fraud detection method that can achieve minute-level or even millisecond-level response speed without losing effectiveness and stability remains a challenging proposition.

To address the issues mentioned above, work must be carried out in the following two aspects: ① Continuing research on distributed machine learning and using technical approaches such as the computational parallel mode, data-parallel mode, and model parallel mode to make full use of existing big data and large models; and ② constructing lightweight networks. Through knowledge distillation, pruning, and other model-compression techniques, models can be made lightweight and customized to facilitate the design of a faster risk detection algorithm.

5.4. Research direction 4: Risk detection methods for low-resource scenarios

The current success of deep learning in many fields is due to the support of large-scale labeled datasets. Most existing models also employ supervised or semi-supervised learning paradigms. However, labeled data in a tax scenario is very difficult to obtain. Even without considering the privacy and security issues associated with tax data, it is impossible to label companies with tax risk behaviors through crowdsourcing and other methods, because this would require a wealth of expert tax-related knowledge and experience. Building tax datasets containing a large amount of label information is accordingly very expensive. However, designing a cognitive inference risk detection method for low-resource scenarios presents the following challenges: ① Firstly, there is limited labeled data available in low-resource scenarios, making it difficult to train and evaluate the performance of the model. ② Second, in tax scenarios, only a very small fraction of enterprises are labeled as risky, leaving a large number of enterprises unlabeled. ③ Third, start-ups have very little transaction information, making it difficult to accurately assess the risks of new enterprises.

Given the above problems, it is necessary to focus on the following four aspects in future work: ① Using active learning, actively selecting the most valuable samples to label, and thereby maximize model benefits with minimized overhead; ② Using unsupervised learning methods such as comparative learning, generative models, and clustering methods to design models for low-resource tax scenarios; ③ Conducting research on semi-supervised methods such as PU learning in order to fully utilize unlabeled samples; and ④ researching technologies such as meta learning, data enhancement, and transfer learning in order to more accurately assess the risks of new enterprises.

6. Conclusions

To accelerate the high-quality development of artificial intelligence in the field of tax risk detection and to better assist national tax authorities in tax risk detection and decision-making, this survey comprehensively reviewed the research progress of tax risk detection at home and abroad for the first time, and summarized the advantages and disadvantages of each method. We also analyzed the limitations of current tax risk detection methods and summarized four research problems—namely, the difficult integration and utilization of fragmented fiscal and tax knowledge, unexplainable risk detection results, the high cost of risk detection algorithms, and the dependence of existing algorithms on label information—and charted the future development direction of tax risk detection from informatization to intellectualization.

Acknowledgments

This research was partially supported by the Key Research and Development Project in Shaanxi Province (2023GXLH-024), the National Natural Science Foundation of China (62250009, 62002282, 62037001, and 62192781).

Compliance with ethics guidelines

Qinghua Zheng, Yiming Xu, Huixiang Liu, Bin Shi, Jiaxiang Wang, and Bo Dong declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Wang D, Huang Y, Cai Z. The State Council Information Office held a press conference on tax and fee reduction to boost confidence in the development [Internet]. Beijing: The State Council Information Office of the People's Republic of China; 2022 Jan 26 [cited 2022 Nov 1]. Available from: http://www.scio.gov.cn/xwfbh/xwbfbh/wqfbh/47673/47802/index.htm. [Chinese].
[2] The tax gap-tax gap estimates for tax years 2014–2016 [Internet]. Washington: Internal Revenue Service; 2022 Oct 28 [cited 2022 Nov 1]. Available from: https://www.irs.gov/newsroom/the-tax-gap.
[3] Androniceanu A, Gherghina R, Ciobănașu M. The interdependence between fiscal public policies and tax evasion. Adm Si Manag Public 2019;32:32–41.
[4] López JJ. A quantitative theory of tax evasion. J Macroecon 2017;53:107–26.
[5] Allingham MG, Sandmo A. Income tax evasion: a theoretical analysis. J Public Econ 1972;1(3–4):323–38.
[6] Zhao Q, Bhowmick SS. Association rule mining: a survey. Report. Singapore: Nanyang Technological University; 2003.
[7] Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining—a general survey and comparison. SIGKDD Explor 2000;2(1):58–64.
[8] Wu RS, Ou CS, Lin H, Chang SI, Yen DC. Using data mining technique to enhance tax evasion detection performance. Expert Syst Appl 2012;39(10):8769–77.
[9] Matos T, de Macedo JAF, Monteiro JM. An empirical method for discovering tax fraudsters: a real case study of Brazilian fiscal evasion. In: Proceedings of the 19th International Database Engineering & Applications Symposium; 2015 Jul 13–15; Yokohama, Japan. New York City: Association for Computing Machinery (ACM); 2015. p. 41–8.
[10] Zhao Z, Jian Z, Gaba GS, Alroobaea R, Masud M, Rubaiee S. An improved association rule mining algorithm for large data. J Intell Syst 2021;30(1):750–62.
[11] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660–74.
[12] Clark LA, Pregibon D. Tree-based models. In: Hastie TJ, editor. Statistical models in S. New York City: Taylor & Francis Group; 2017. p. 377–419.
[13] 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; 1999 Aug 30–Sep 1; Florence, Italy. Berlin: Springer; 1999. p. 369–76.
[14] Mittal S, Reich O, Mahajan A. Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies; 2018 Jun 20–22; Menlo Park and San Jose, CA, USA. New York City: Association for Computing Machinery (ACM); 2018. p. 1–11.
[15] Yao J, Zhang J, Wang L. A financial statement fraud detection model based on hybrid data mining methods. In: Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD); 2018 May 26–28; Chengdu, China. New York City: IEEE; 2018. p. 57–61.
[16] Wu C, Luo J. Automatic recognition of tax evasion behavior based on random forest. Software Guide 2018;017(008):13–6.
[17] An B, Suh Y. Identifying financial statement fraud with decision rules obtained from modified random forest. Data Technol Appl 2020;54(2):235–55.
[18] Ji YL, Wang WQ. The stock of research on accurate identification of tax risk under the background of big data technology–based on machine learning. Public Finance Res 2020;451(09):121–31 [Chinese].
[19] Andrade JPA, Paulucio LS, Paixao TM, Paixao TM, Berriel RF, Carneiro TCJ, et al. A machine learning-based system for financial fraud detection. In: Proceedings of the 18th National Meeting on Artificial and Computational Intelligence (ENIAC 2021); 2021 Nov 29–Dec 3; online. São Leopoldo: Sociedade Brasileira de Computação (SBC); 2021. p. 165–76.
[20] Xavier OC, Pires SR, Marques TC, Soares AS. Tax evasion identification using open data and artificial intelligence. Rev Adm Pública 2022;56(3):426–40.
[21] Agarwal A, Tan YS, Ronen O, Singh C, Yu B. Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based models. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. New York City: MLResearch Press; 2022. p. 111–35.
[22] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565–7.
[23] Pisner DA, Schnyer DM. Support vector machine. In: Mechelli A, Vieira S, editors. Machine learning. Cambridge: Academic Press; 2020. p. 101–21.
[24] Wang S, Li A. Fraud detection in tax declaration based on SVM. Comput Eng 2006;.
[25] Liu H, Yu X, Wan W, Ma X. A tax assessment model based on rough set theory and SVM algorithms. Comput Simu 2009;26(12):253–6 [Chinese].
[26] Xia H, Li R. Cases-choice in tax declaration model based on SVM and SOM. Sci Technol Eng 2009;009(014):4027–31 [Chinese].
[27] Junqué de Fortuny E, Stankova M, Moeyersoms J, Minnaert B, Provost FJ, Martens D. Corporate residence fraud detection. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2014 Aug 24–27; New York City, NY, USA. New York City: Association for Computing Machinery (ACM); 2014. p. 1650–9.
[28] Rad MS, Shahbahrami A. Detecting high risk taxpayers using data mining techniques. In: Proceedings of the 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS 2016); 2016 Dec 14–15; Tehran, Iran. New York City: IEEE; 2016 .p. 1–5.
[29] Zhang X. Early warning and investigation countermeasures of crime of issuing false invoice [dissertation]. Beijing: People's Public Security University of China; 2020 [Chinese].
[30] Cervantes J, Garcia-Lamont F, Rodriguez L, López A, Castilla JR, Trueba A. PSO-based method for SVM classification on skewed data sets. Neurocomputing 2017;228:187–97.
[31] Rish I. An empirical study of the Naive Bayes classifier. In: Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence; 2001 Aug 4–6; Washington, DC, USA. Berlin: Springer; 2001. p. 41–6.
[32] Leung KM. Naive Bayesian classifier. Hong Kong: Polytechnic University Department of Computer Science/Finance and Risk Engineering; 2007.
[33] Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. Expert Syst Appl 2007;32(4):995–1003.
[34] Kang Z, Yu Y. Study on tax evaluation model based Bayesian classification. Econ Probl 2009;6:124–6. Chinese.
[35] Zhang K, Wu D, Li A, Song BW. Fraud detection in tax declaration based on Bayesian classifier. Comput Simu 2010;27(009):306–10 [Chinese].
[36] Lenz HJ. Tax fraud and investigation procedures-everybody, every where, every time. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016); 2016 Feb 19–21; Rome, Italy. Trier: The DBLP Computer Science Bibliography; 2016. p. 3–13.
[37] Zaidi NA, Cerquides J, Carman MJ, Webb GI. Alleviating Naive Bayes attribute independence assumption by attribute weighting. J Mach Learn Res 2013;14(60):1947–88.
[38] Kleinbaum DG, Klein M. Logistic regression: a self-learning text. 2nd ed. New York City: Springer-Verlag; 2002.
[39] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken: John Wiley & Sons; 2013.
[40] Qi X. The research on the tax inspection methods about identifying tax evasion [dissertation]. Changchun: Jilin University; 2010 [Chinese].
[41] Wang Y, Li Q, Qi X. Research on the tax inspection selection scheme model based on the Logistic regression. Econ Res Guide 2012;35(2):96–7. Chinese.
[42] Su Y. Research on tax inspection case selection based on logistic regression model [dissertation]. Guangzhou: Sun Yat-sen University; 2011. Chinese.
[43] Yuan Y. Research on the model of tax inspection and case selection in H city based on logistic regression to identify enterprise tax evasion [dissertation]. Hohhot: Inner Mongolia University; 2019 [Chinese].
[44] Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting. Nat Methods 2016;13(9):703–5.
[45] Frades I, Matthiesen R. Overview on techniques in cluster analysis. Bioinformatics methods in clinical research 2010:81–107.
[46] Duran BS, Odell PL. Cluster analysis: a survey. Berlin: Springer Science & Business Media; 2013.
[47] Denny, Williams G J, Christen P. Exploratory multilevel hot spot analysis: Australian taxation office case study. In: Proceedings of the 6th Australasian Conference on Data Mining and Analytics-Volume 70; 2007 Dec 3–4; Queensland, QLD, Australia. New York City: Association for Computing Machinery (ACM); 2007. p. 77–84.

[48] Liu X, Pan D, Chen S. Application of hierarchical clustering in tax inspection case-selecting. In: Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering; 2010 Dec 10–12; Wuhan, China. New York City: IEEE; 2010. p. 1–4.
[49] Liu B, Xu G, Xu Q, Zhang N. Outlier detection data mining of tax based on cluster. Phys Procedia 2012;33:1689–94.
[50] Assylbekov Z, Melnykov I, Bekishev R, Baltabayeva A, Bissengaliyeva D, Mamlin E. Detecting value-added tax evasion by business entities of Kazakhstan. In: Czarnowski I, Caballero A, Howlett R, Jain L, editors. Proceedings of the International Conference on Intelligent Decision Technologies; 2016 Jun 15–17; Puerto de la Cruz, Spain. Berlin: Springer, Cham; 2016. p. 37–49.
[51] de Roux D, Perez B, Moreno A, del Pilar VM, Figueroa C. In: Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. London, UK. New York City: Association for Computing Machinery (ACM); 2018. p. 215–22.
[52] Xia H, Cheng P, Zhang L. Tax risk identification based on improved K-means clustering algorithm under big data. Fin Accou Mon 2019;21:143–6. Chinese.
[53] Ben-David S, Haghtalab N. In: Clustering in the presence of background noise. Beijing, China. New York City: MLResearch Press; 2014. p. 280–8.
[54] Guo X, Li S. In: Distributed *k*-clustering for data with heavy noise. Montréal, QC, Canada. New York City: Association for Computing Machinery (ACM); 2018. p. 7849–57.
[55] Bishop CM. Neural networks and their applications. Rev Sci Instrum 1994;65 (6):1803–32.
[56] Khan S, Rahmani H, Shah SAA, Bennamoun M. A guide to convolutional neural networks for computer vision. Berlin: Springer; 2018.
[57] Sinkov A, Asyaev G, Mursalimov A, Nikolskaya K. In: Neural networks in data mining. Chelyabinsk, Russia. New York City: IEEE; 2016. p. 1–5.
[58] Zhang J, Zong C. Deep neural networks in machine translation: an overview. IEEE Intell Syst 2015;30(5):16–25.
[59] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: a survey. Heliyon 2018;4 (11):e00938.
[60] Li S, Xiao X. Application of tax payment evaluation based on fuzzy neural network. Comput Simu 2012;29(01):352–5 [Chinese].
[61] Lin CC, Chiu AA, Huang SY, Yen DC. Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgments. Knowl Base Syst 2015;89:459–70.
[62] Assylbekov Z, Melnykov I, Bekishev R, Baltabayeva A, Bissengaliyeva D, Mamlin E. Detecting value-added tax evasion by business entities of Kazakhstan. In: Proceedings of the International Conference on Intelligent Decision Technologies; 2016 Jun 15–17; Puerto de la Cruz, Spain. Berlin: Springer, Cham; 2016. p. 37–49.
[63] Pérez López C, Delgado Rodríguez MJ, de Lucas SS. Tax fraud detection through neural networks: an application using a sample of personal income taxpayers. Future Internet 2019;11(4):86.
[64] Zhang L, Nan X, Huang E, Liu S. Detecting transaction-based tax evasion activities on social media platforms using multi-modal deep neural networks. 2020. arXiv:2007.13525.
[65] Chen H, Gong L, Cheng L, You Z. Tax risk assessment model of large enterprises based on multilayer perceptron. Appl Res Comput 2020;37 (S2):41–3+6. Chinese.
[66] Zhang L, Nan X, Huang E, Liu S. In: Social E-commerce tax evasion detection using multi-modal deep neural networks. Gold Coast, QLD, Australia. New York City: IEEE; 2021. p. 1–6.
[67] Murorunkwere BF, Tuyishimire O, Haughton D, Nzabanita J. Fraud detection using neural networks: a case study of income tax. Future Internet 2022;14 (6):168.
[68] Mojahedi H, Babazadeh Sangar A, Masdari M. Towards tax evasion detection using improved particle swarm optimization algorithm. Math Probl Eng 2022;2022:1027518.
[69] Alsadhan NA. Value-added tax fraud detection and anomaly feature selection using sectorial autoencoders. In: Proceedings of the Data Analytics and Management (ICDAM 2022); 2022 Jun 25–26; Jelenia Góra, Poland. Singapore: Springer; 2022. p. 323–31.
[70] Fan FL, Xiong J, Li M, Wang G. On interpretability of artificial neural networks: a survey. IEEE Trans Radiat Plasma Med Sci 2021;5(6):741–60.
[71] Kar K, Kornblith S, Fedorenko E. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. Nat Mach Intell 2022;4 (12):1–3.
[72] Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018;106:249–59.
[73] Trentin E, Gori M. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing 2001;37(1–4):91–126.
[74] Ravisankar P, Ravi V, Rao GR, Bose I. Detection of financial statement fraud and feature selection using data mining techniques. Decis Support Syst 2011;50(2):491–500.
[75] Zheng M. Research on tax data mining based on SAS system [dissertation]. Zhengzhou: Zhengzhou University; 2012 [Chinese].
[76] González PC, Velásquez JD. Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Syst Appl 2013;40 (5):1427–36.
[77] Song XP, Hu ZH, Du JG, Sheng ZH. Application of machine learning methods to risk assessment of financial statement fraud: evidence from China. J Forecast 2014;33(8):611–26.
[78] Rahimikia E, Mohammadi S, Rahmani T, Ghazanfari M. Detecting corporate tax evasion using a hybrid intelligent system: a case study of Iran. Int J Account Inf Syst 2017;25:1–17.
[79] Wu Y, Zheng Q, Gao Y, Dong B, Wei R, Zhang F, et al. In: TEDM-PU: a tax evasion detection method based on positive and unlabeled learning. Los Angeles, CA, USA. New York City: IEEE; 2019. p. 1681–6.
[80] Javadian Kootanaee A, Poor Aghajan Sarhamami AA, Hosseini SM. A model for identification tax fraud based on improved id3 decision tree algorithm and multilayer perceptron neural network. Manag Account 2020; 13(46):53–70.
[81] Rahman RA, Masrom S, Omar N, Zakaria M. An application of machine learning on corporate tax avoidance detection model. IAES Int J Artif Intell 2020;9(4):721.
[82] Mekonnen E. Data mining for detection of tax evasion: the case of tax payers in Addis Ababa [dissertation]. London: St. Mary's University; 2021.
[83] Savić M, Atanasijević J, Jakovetić D, Krejić N. Tax evasion risk management using a hybrid unsupervised outlier detection method. Expert Syst Appl 2022;193:116409.
[84] Baghdasaryan V, Davtyan H, Sarikyan A, Navasardyan Z. Improving tax audit efficiency using machine learning: the role of taxpayer's network data in fraud detection. Appl Artif Intell 2022;36(1):2012002.
[85] Schunck R. Within and between estimates in random-effects models: advantages and drawbacks of correlated random effects and hybrid models. Stata J 2013;13(1):65–76.
[86] Zhu X, Yan Z, Ruan J, Zheng Q, Dong B. In: IRTED-TL: an inter-region tax evasion detection method based on transfer learning. New York City, NY, USA. New York City: IEEE; 2018. p. 1224–35.
[87] Wei R, Dong B, Zheng Q, Zhu X, Ruan J, He H. In: Unsupervised conditional adversarial networks for tax evasion detection. Los Angeles, CA, USA. New York City: IEEE; 2019. p. 1675–80.
[88] Zhang F, Shi B, Dong B, Zheng Q, Ji X. In: TTED-PU: a transferable tax evasion detection method based on positive and unlabeled learning. Madrid, Spain. New York City: IEEE; 2020. p. 207–16.
[89] Wang J, Chen Y. Safe and robust transfer learning. Singapore: Springer; 2022.
[90] Nam J, Pan SJ, Kim S. In: Transfer defect learning. San Francisco, CA, USA. New York City: IEEE; 2013. p. 382–91.
[91] Li Y. Deep reinforcement learning: an overview. 2017. arXiv:1701.07274.
[92] Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge: MIT Press; 2018.
[93] François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J. An introduction to deep reinforcement learning. Hanover: Now Foundations and Trends®. Mach Learn 2018.
[94] Abe N, Melville P, Pendus C, Reddy C, Jensen D, Thomas V. In: Optimizing debt collections using constrained reinforcement learning. Washington, DC, USA. New York City: Association for Computing Machinery (ACM); 2010. p. 75–84.
[95] Goumagias ND, Hristu-Varsakelis D, Assael YM. Using deep Q-learning to understand the tax evasion behavior of risk-averse firms. Expert Syst Appl 2018;101:258–70.
[96] Bonnet C, Caron P, Barrett T, Davies I, Laterre A. One step at a time: pros and cons of multi-step meta-gradient reinforcement learning. 2021. arXiv:2111.00206.
[97] Jitani A, Mahajan A, Zhu Z, Abou-Zeid H, Fapi ET, Purmehdi H. Structure-aware reinforcement learning for node-overload protection in mobile edge computing. IEEE Trans Cogn Commun Netw 2022;8(4):1881–97.
[98] Bäck T, Schwefel HP. An overview of evolutionary algorithms for parameter optimization. Evol Comput 1993;1(1):1–23.
[99] Bartz-Beielstein T, Branke J, Mehnen J, Mersmann O. Evolutionary algorithms. Wiley Interdiscip Rev Data Min Knowl Discov 2014;4(3):178–95.
[100] Alden ME, Bryan DM, Lessley BJ, Tripathy A. Detection of financial statement fraud using evolutionary algorithms. J Emerg Technol Account 2012;9 (1):71–94.
[101] Warner G, Wijesinghe S, Marques U, Badar O, Rosen J, Hemberg E, et al. Modeling tax evasion with genetic algorithms. Econ Gov 2015;16(2):165–78.
[102] Hemberg E, Rosen J, Warner G, Wijesinghe S, O'Reilly UM. Tax non-compliance detection using co-evolution of tax evasion risk and audit likelihood. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law; 2015 Jun 8–12; San Diego, CA, USA. New York City: Association for Computing Machinery (ACM); 2015. p. 79–88.
[103] Hemberg E, Rosen J, Warner G, Wijesinghe S, O'Reilly UM. Detecting tax evasion: a co-evolutionary approach. Artif Intell Law 2016;24(2):149–82.
[104] Karafotias G, Hoogendoorn M, Eiben ÁE. Parameter control in evolutionary algorithms: trends and challenges. IEEE Trans Evol Comput 2014;19 (2):167–87.
[105] Lobo FG, Lima C, Michalewicz Z. Parameter setting in evolutionary algorithms. Berlin: Springer Science & Business Media; 2007.
[106] Sipper M, Fu W, Ahuja K, Moore JH. Investigating the parameter space of evolutionary algorithms. BioData Min 2018;11(1):2.
[107] Gilbert N, Terna P. How to build and use agent-based models in social science. Mind & Society 2000;1:57–72.
[108] Samanidou E, Zschischang E, Stauffer D, Lux T. Agent-based models of financial markets. Rep Prog Phys 2007;70(3):409–50.
[109] Gilbert N. Agent-based models. Newbury Park: SAGE Publications; 2019.
[110] Antunes L, Balsa J, Coelho H. Agents that collude to evade taxes. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems; 2007 May 14–18; Honolulu, HI, USA. New York City: Association for Computing Machinery (ACM); 2007. p. 1–3.

[111] Lima FWS. Tax evasion and nonequilibrium model on apollonian networks. Int J Mod Phys C 2012;23(11):1250079.
[112] Llacer T, Miguel FJ, Noguera JA. Tapia E An agent-based model of tax compliance: an application to the Spanish case. Advances in Complex Systems 2013;16(04n05):1350007.
[113] Noguera JA, Quesada FJM, Tapia E, Llàcer T. Tax compliance, rational choice, and social influence: an agent-based model. Rev Fr Sociol 2014;55 (4):765–804.
[114] Andrei AL, Comer K, Koehler M. An agent-based model of network effects on tax compliance and evasion. J Econ Psychol 2014;40:119–33.
[115] Bloomquist KM. A comparison of agent-based models of income tax evasion. Soc Sci Comput Rev 2006;24(4):411–25.
[116] Manzo G, Matthews T. Potentialities and limitations of agent-based simulations. Rev Fr Sociol 2014;55(4):653–88.
[117] McDonald GW, Osgood ND. Agent-based modeling and its tradeoffs: an introduction & examples. 2023. arXiv:2304.08497.
[118] Fan W. Graph pattern matching revised for social network analysis. In: Proceedings of the 15th International Conference on Database Theory; 2012 Mar 26–29; Berlin, Germany. New York City: Association for Computing Machinery (ACM); 2012. p. 8–21.
[119] Ma S, Cao Y, Fan W, Huai JP, Wo T. Strong simulation: capturing topology in graph pattern matching. ACM Transactions on Database Systems 2014;39 (1):1–46.
[120] Tian F, Lan T, Chao KM, Godwin N, Zheng Q, Shah N, et al. Mining suspicious tax evasion groups in big data. IEEE Trans Knowl Data Eng 2016;28 (10):2651–64.
[121] Wei W, Yan Z, Ruan J, Zheng Q, Dong B. Mining suspicious tax evasion groups in a corporate governance network. In: Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing; 2017 Aug 21–23; Helsinki, Finland. Berlin: Springer, Cham; 2017. p. 465–75.
[122] Liu L. Methods of detect falsely making out specialized invoices behavior based on directed graph [dissertation]. Xi'an: Xi'an University of Science and Technology; 2017. Chinese.
[123] Ruan J, Yan Z, Dong B, Zheng Q, Qian B. Identifying suspicious groups of affiliated-transaction-based tax evasion in big data. Inf Sci 2019;477:508–32.
[124] Mathews J, Mehta P, Babu S. Link prediction techniques to handle tax evasion. In: Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD); 2021 Jan 2–4; online. New York City: Association for Computing Machinery (ACM); 2021. p. 307–15.
[125] Rocha-Salazar JJ, Segovia-Vargas MJ, Camacho-Miñano MM. Detection of shell companies in financial institutions using dynamic social network. Expert Syst Appl 2022;207:117981.
[126] Chen T, Tsourakakis C. Antibenford subgraphs: unsupervised anomaly detection in financial networks. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2022 Aug 14–18; Washington, DC, USA. New York City: Association for Computing Machinery (ACM); 2022. p. 2762–70.
[127] Fan W, Li J, Ma S, Tang N, Wu Y, Wu Y. Graph pattern matching: from intractable to polynomial time. Proceedings VLDB Endowment 2010;3(1–2):264–75.
[128] Ma S, Cao Y, Huai J, Wu T. Distributed graph pattern matching. In: Proceedings of the 21st International Conference on World Wide Web Conference; 2012 Apr 16–20; Lyon, France. New York City: Association for Computing Machinery (ACM); 2012. p. 949–58.
[129] Bouhenni S, Yahiaoui S, Nouali-Taboudjemat N, Kheddouci H. A survey on distributed graph pattern matching in massive graphs. ACM Comput Surv 2021;54(2):1–35.
[130] Chen F, Wang YC, Wang B, Kuo CCJ. Graph representation learning: a survey. APSIPA Trans Signal Inf Process 2020;9(1):e15.
[131] Khoshraftar S, An A. A survey on graph representation learning methods. 2022. arXiv:2204.01855.
[132] Matos T, de Macêdo JAF, Monteiro JM, Lettich F. An accurate tax fraud classifier with feature selection based on complex network node centrality measure. In: Proceedings of the 19th International Conference on Enterprise Information Systems; 2017 Apr 26–29; Porto, Portugal. Berlin: Springer; 2017.p. 145–51.
[133] Wu Y, Dong B, Zheng Q, Wei R, Wang Z, Li X. A novel tax evasion detection framework via fused transaction network representation. In: Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC); 2020 Jul 13–17; Madrid, Spain. New York City: IEEE; 2020. p. 235–44.
[134] Mi L, Dong B, Shi B, Zheng Q. A tax evasion detection method based on positive and unlabeled learning with network embedding features. In: Proceedings of the International Conference on Neural Information Processing; 2020 Nov 18–22; Bangkok, Thailand. Berlin: Springer; 2020. p. 140–51.
[135] An J, Zheng Q, Wei R, Dong B, Li X. NEUD-TRI: network embedding based on upstream and downstream for transaction risk identification. In: Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC); 2020 Jul 13–17; Madrid, Spain. New York City: IEEE; 2020. p. 277–86.
[136] Wang Y, Zheng Q, Ruan J, Gao Y, Chen Y, Li X, et al. T-EGAT: a temporal edge enhanced graph attention network for tax evasion detection. In: Proceedings of the 2020 IEEE International Conference on Big Data (Big Data); 2020 Dec 10–13; Atlanta, GA, USA. New York City: IEEE; 2020. p. 1410–5.
[137] Gao Y, Shi B, Dong B, Wang Y, Mi L, Zheng Q. Tax evasion detection with FBNE-PU algorithm based on PnCGCN and PU learning. IEEE Trans Knowl Data Eng 2021;35(1):931–44.
[138] Shi B, Dong B, Xu Y, Wang J, Wang Y, Zheng Q. An edge feature aware heterogeneous graph neural network model to support tax evasion detection. Expert Syst Appl 2023;213:118903.
[139] Gogoglou A, Bruss CB, Hines KE. On the interpretability and evaluation of graph representation learning. 2019. arXiv:1910.03081.
[140] Leite RA, Gschwandtner T, Miksch S, Gstrein E, Kuntner J. Visual analytics for event detection: focusing on fraud. Vis Inform 2018;2(4):198–212.
[141] Yuan J, Chen C, Yang W, Liu M, Xia J, Liu S. A survey of visual analytics techniques for machine learning. Comput Vis Media (Beijing) 2021;7 (1):3–36.
[142] Liu D, Alnegheimish S, Zytek A, Veeramachaneni K. MTV: visual analytics for detecting, investigating, and annotating anomalies in multivariate time series. Proc ACM Hum Comput Interact 2022;6(CSCW1):103.
[143] Didimo W, Liotta G, Montecchiani F, Palladino P. An advanced network visualization system for financial crime detection. In: Proceedings of the 2011 IEEE Pacific Visualization Symposium; 2011 Mar 1–4; Hong Kong, China. New York City: IEEE; 2011. p. 203–10.
[144] Tselykh A, Knyazeva M, Popkova E, Durfee A, Tselykh A. An attributed graph mining approach to detect transfer pricing fraud. In: Proceedings of the 9th International Conference on Security of Information and Networks; 2016 Jul 20–22; Newark, NJ, USA. New York City: Association for Computing Machinery (ACM); 2016. p. 72–5.
[145] Didimo W, Giamminonni L, Liotta G, Montecchiani F, Pagliuca D. A visual analytics system to support tax evasion discovery. Decis Support Syst 2018;110:71–83.
[146] Zheng Q, Lin Y, He H, Ruan J, Dong B. ATTENet: detecting and explaining suspicious tax evasion groups. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence; 2019 Aug 10–16; Macao, China. Washington: AAAI Press; 2019. p. 6584–6.
[147] Dai H, Dai B, Song L. Discriminative embeddings of latent variable models for structured data. In: Proceedings of the International Conference on Machine Learning; 2016 Jun 19–24; New York City, NY, USA. New York City: Association for Computing Machinery (ACM); 2016. p. 2702–11.
[148] Yu H, He H, Zheng Q, Dong B. TaxVis: a visual system for detecting tax evasion group. In: Proceedings of the World Wide Web Conference; 2019 May 13–17; San Francisco, CA, USA. New York City: Association for Computing Machinery (ACM); 2019. p. 3610–4.
[149] Grover A, Leskovec J. Node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 6–11; San Francisco, CA, USA. New York City: Association for Computing Machinery (ACM); 2016. p. 855–64.
[150] Didimo W, Grilli L, Liotta G, Menconi L, Montecchiani F, Pagliuca D. Combining network visualization and data mining for tax risk assessment. IEEE Access 2020;8:16073–86.
[151] Zha Z. TaxAA: a reliable tax auditor assistant for exploring suspicious transactions. In: Proceedings of the Web Conference 2020; 2020 Apr 20–4; Taipei, China. New York City: Association for Computing Machinery (ACM); 2020. p. 240–4.
[152] Lin Y, Wong K, Wang Y, Zhang R, Dong B, Qu H, et al. Taxthemis: interactive mining and exploration of suspicious tax evasion groups. IEEE Trans Vis Comput Graph 2021;27(2):849–59.
[153] Nussbaumer A, Verbert K, Hillemann EC, Bedek MA, Albert D. A framework for cognitive bias detection and feedback in a visual analytics environment. In: Proceedings of the 2016 European Intelligence and Security Informatics Conference (EISIC 2016); 2016 Aug 16–19; Uppsala, Sweden. New York City: IEEE; 2016. p. 148–51.
[154] Wall E, Blaha LM, Franklin L, Endert A. Warning, bias may occur: a proposed approach to detecting cognitive bias in interactive visual analytics. In: Proceedings of the 2017 IEEE Conference On Visual Analytics Science And Technology (VAST 2017); 2017 Oct 3–6; Phoenix, AZ, USA. New York City: IEEE; 2017. p. 104–15.
[155] Wall E. Detecting and mitigating human bias in visual analytics [dissertation]. Atlanta: Georgia Institute of Technology; 2020.
[156] Zheng Q. 2019 big data knowledge engineering and application. J Comput Res Dev 2019;56(12):2519–20.
[157] Wu F, Han Y, Li X, Zheng QH. Chen XL Reasoning in artificial intelligence: advances and challenges. Bull Natl Nat Sci Found Chin 2018;32(3):262–5 [Chinese].
[158] Zhuang Y, Wu F, Chen C, Pan Y. Challenges and opportunities: from big data to knowledge in AI 2.0. Front Inf Technol Electron Eng 2017;18(1):3–14.
[159] Yang Y, Zhuang Y, Pan Y. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. Front Inf Technol Electron Eng 2021;22(12):1551–8.
[160] Zheng Q, Liu J, Zeng H, Guo Z, Wu B, Wei B. Knowledge forest: a novel model to organize knowledge fragments. Sci China Inf Sci 2021;64(7):179103.