Expert Systems with Applications 39 (2012) 8769–8777

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Using data mining technique to enhance tax evasion detection performance

Roung-Shiunn Wu [a], C.S. Ou [b], Hui-ying Lin [b], She-I Chang [b], David C. Yen [c,*]

[a] Department of Information Management, National Chung Cheng University, 168 University Rd., Min-Hsiung, Chia-Yi, Taiwan, ROC
[b] Department of Accounting and Information Technology, National Chung Cheng University, 168 University Rd., Min-Hsiung, Chia-Yi, Taiwan, ROC
[c] Department of DSC and MIS, Miami University, 2042C, FSB, Miami University, Oxford, OH 45056, USA

ARTICLE INFO

Keywords:
Data mining
Value-added tax
Tax evasion
Association rule

ABSTRACT

Currently, tax authorities face the challenge of identifying and collecting from businesses that have successfully evaded paying the proper taxes. In solving the problem of tax evaders, tax authorities are equipped with limited resources and traditional tax auditing strategies that are time-consuming and tedious. These continued practices have resulted in the loss of a substantial amount of tax revenue for the government. The objective of the current study is to apply a data mining technique to enhance tax evasion detection performance. Using a data mining technique, a screening framework is developed to filter possible non-compliant value-added tax (VAT) reports that may be subject to further auditing. The results show that the proposed data mining technique truly enhances the detection of tax evasion, and therefore can be employed to effectively reduce or minimize losses from VAT evasion.

1. Introduction

Tax revenue is one of the most necessary financial resources of a government for accomplishing specific goals. However, some businesses often attempt to evade their payment of correct taxes. Consequently, tax evasion creates a critical impact on the budgetary income of these businesses and of the government. These businesses incur additional social costs because they spend their valuable resources in finding means to evade taxes, rather than focus on their operations. On the side of the government, tax authorities have to bear the costs of the detection and prevention of illegal tax evasion activities. As a result, effective ways to detect related tax evasion activities have always been an important and challenging issue for tax authorities in any country.

If the government cannot effectively detect illegal tax evasion activities, public investment would be negatively affected due to the budgetary shortage resulting from the loss of tax revenues. VAT (value-added tax) evasion is one of the important issues for many tax authorities. Gebauer, Nam, and Parsche (2007) report that, based on German data, the VAT revenue gap derived from the comparison of the quantified, hypothetical, and the actual collected revenues increased from 5.1% in 1995 to 7.5% in 1996. They also estimate that VAT revenue losses were approximately EURO 18 billion in 2001 for Germany alone.

Gebauer et al. (2007) also suggest that VAT evasion not only leads to significant revenue losses, but also to a considerable increase in administrative costs used to detect the illegal tax evasion activities. In addition to the decreased tax revenue and increased administrative costs, VAT evasion also has a significant negative side effect on the collection of income taxes from businesses. This occurs because VAT evasion, which implies an indirectly underreported taxable income from business, is often directly accompanied by underreported sales revenues.

In order to realize the benefits of spending valuable, albeit limited, resources to detect VAT evasion, tax authorities need to deploy their resources wisely. As such, tax authorities have often relied on the sampling method and the personal judgment of tax auditors in selecting suspicious tax reports to audit for potential tax evasion activities. Thus, the purpose of the current study is to determine a more scientific approach to improve tax auditor's productivity and performance in handling the detection tasks of VAT tax evasion.

All over the world, tax authorities are under increasing pressure to locate underreporting taxpayers, collect additional tax revenues, and predict the irregular behavior of non-paying taxpayers. Without the assistance of information technology tools, most tax authorities need to pull in tax data from a variety of independent sources or perform data matching and checking with other sources to find cases of non-compliance. As a result, tax evasion detection performance has been rather limited.

Business intelligence (BI) in general, and data mining in particular, may be effective tools for enhancing the efficiency and effectiveness of the detection of illegal tax evasion (Fadairo, Williams, Trotman, & Onyekelu-Eze, 2008). In the US, Texas was one of the

* Corresponding author. Tel.: +1 513 529 4827; fax: +1 513 529 9689.
E-mail addresses: roungwu@ccu.edu.tw (R.-S. Wu), actcso@gmail.com (C.S. Ou), enya@email.ntact.gov.tw (H.-y. Lin), actsic@ccu.edu.tw (S.-I. Chang), yendc@muohio.edu (D.C. Yen).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.204

first states to apply data mining techniques for detecting suspicious tax evasion reports and thereby recoup unpaid taxes (Hoover, 2009). Songini (2004) reports that Texas uses a BI system that is able to flag a situation in which a business is suspected to be evading taxes. This suspicious tax report is referred to an audit staff for further investigation. Since the introduction and application of the BI system, USD 362 million of tax losses have already been recovered. The tax authority in Texas has also committed strongly to data mining for spotting suspicious tax reports. As cited by Songini (2004), Lisa McCormack, a manager in the tax audit division in Austin, Texas, claims, "We only audit 1% of the taxpayers... We have to try and figure out how to make the best use of the [government's investigative] resources."

The current study intends to utilize data mining as a tool to enhance tax evasion detection performance. Data mining is a methodology used to discover hidden information from rough data (Fayyad, Piatesky-Shapiro, & Padhraic, 1996; Yoon, 1999). It can be applied in the process of decision support, prediction, forecasting, and estimation (Liao, 2003). Moreover, data mining techniques are able to efficiently handle a large number of records and data (Ravisankar, Ravi, Raghava Rao, & Bose, 2011). Compared to general statistics, data mining is able to identify certain patterns and match specific data via efficient computing technology. In other words, the interpretation of data allows flexibility (Liao, 2003).

This study employs the association rule of the data mining technique to the VAT database to uncover patterns and relationships among attributes that are useful for identifying problematic tax evasion reports. In this research, a screening model will be developed based on specific patterns or rules discovered from identified VAT evasion tax reports. This screening model is utilized to select the cases that are suspected to be non-compliant VAT reports for further auditing checks. In other words, the goal of using data mining as a technique in detecting VAT evasion enhances the tax auditor's productivity in recovering tax revenue losses.

The current paper is organized as follows. After the introductory section, Section 2 provides a literature review. Section 3 illustrates the proposed framework. Sections 4 and 5 discuss the design and development of the screening model and the experimental results, respectively. Finally, Section 6 provides the conclusion, including the limitations of this study and future implications.

2. Literature review

2.1. Value-added tax evasion detection in Taiwan

Keen and Lockwood (2010) in exploring the causes and consequences of the remarkable worldwide attention given to VAT in recent years, find that more than 130 countries have implemented the VAT scheme. In addition, VAT has raised 20% or more of all tax revenues in those countries. Their estimated figures also suggest that the adoption of VAT contributes positively in the establishment of an effective tax system for most countries under this study. They argue that, "By any standards, the rise of the VAT has been the most significant development in tax policy and administration of recent decades."

The goal of VAT is to collect taxes on the difference between receipts from sales and expenditures on purchases for each business transaction. According to the Department of Investment Services in Taiwan (2010):

> "All sales of goods and provision of services in Taiwan, as well as all imports of goods into Taiwan, are subject to Business Tax (Sales Tax)... The sale of goods is defined as the transfer of goods to another entity for consideration in Taiwan... The sale of services is defined as the supply of services to others or the provision of goods for the use, production of earnings by others for a consideration... The import of goods includes all goods that are imported into Taiwan from a foreign country."

In addition, tax authorities in Taiwan use Input Documentary Evidence (IDE) as a receipt from sales and Output Documentary Evidence (ODE) as an expenditure on purchase. In a business transaction, the buyer obtains the IDE and the seller possesses the ODE. According to statistics from the Ministry of Finance (2011), VAT is an important tax income among overall taxes, ranking only second to income tax from 2003 to 2010. However, according to Huang and Lin (2009), in 1999–2001, more than 35% selected cases in Taiwan involved VAT evasion. Clearly, tax evasion is a serious problem in Taiwan.

One of the techniques that tax authorities traditionally use to detect VAT evasion is to conduct a cross-matching of the IDE and the ODE. Due to upstream to downstream interdependence, the auditing of a seller's VAT report is usually useful for auditing the validity of a buyer's VAT reports in detecting VAT evasion. Once a mismatch between the costs in IDE and the sales in ODE is detected, penalties can be imposed on the guilty taxpayers.

The current cross-matching processes are done manually by tax auditing staff with the help of simple computer software tools. This task can be very tedious and time-consuming. In addition, the effectiveness of this auditing task heavily relies on the experience and skills of the tax auditing staff. Dealing with the huge volume of IDEs and ODEs and processing all VAT reports in an efficient and effective manner might be impossible. In reality, tax authorities are able to screen only a small percentage of tax reports for further auditing because of limited staff resources. Consequently, a systematic approach to identify the VAT reports with high tax-evasion potential among all tax reports becomes a critical and timely necessity for tax authorities.

Gebauer et al. (2007) have examined VAT evasion caused by carousel fraud. Their results indicate that firms in the EU repeatedly carry out cross-border transactions. Carousel fraud actually leads to the fraudulent retention of revenues. Their study examines three reform models of the VAT system in Germany for detecting VAT evasion. They argue that, under the new system, many small cases of fraud can cause an enormous amount of tax losses, which may actually generate a more severe damage than a few large cases. In addition, the administrative overhead, the associated personnel, and equipment costs may be unintentionally underestimated. Therefore, they suggest that, before a radical change in the VAT system is made, tax authorities should run available options to make the present VAT system more effective in the prevention and detection of tax evasion.

Some taxpayers intentionally become involved with some optimal degree of tax evasion. Gupta (2008) analyzes the relationship between tax evasion behavior and taxation policy. He further suggests that, from a policy perspective, increasing the penalty rates may turn to be the best way to reduce tax evasion when taxpayers face budgetary income pressures.

2.2. Data warehouse technologies

More and more companies are using BI tools to analyze sales and other related transactional data to detect fraud. Software companies, such as SAS Institute Inc., SPSS Inc., NCR's Teradata, and IBM's Cognos Business Intelligence all provide fraud-detection oriented BI tools. Some US government agencies are also employing BI tools to detect tax evasion. According to Lisa McCormack, a manager in the audit division in Austin, the comptroller's office in Texas relies on a data warehouse tool to check for sales tax compliance.

A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile dataset (Inmon, 1996). Data warehousing is utilized to collect data from different sources that can then be or-

ganized into semantic and integrated data storage (Nycz & Smok, 2006). Data warehousing is used to support structured and specific queries. The results of queries are mostly analytical reports that can be employed to support decision making.

The objective of data mining is to identify potential useful correlations and patterns in the existing data sources. The most fertile data source is the corporate or organization data warehouse, which integrates all the information from multiple operational sectors. The information stored in the data warehouse can capture many aspects of the business process across different functional areas. Some data are analyzed directly from the data warehouse through the utilization of online analytical processing (OLAP) capability. OLAP tools are able to perform reporting functions where the criteria for either aggregating or summarizing are explicit. Through data mining, more patterns can be discovered to generate important insights and provide additional information. In particular, association rule mining is one of the successful data mining methods for finding hidden patterns and correlations.

Multidimensional modeling, one of the most used techniques to conceptualize the data in data warehouses, organizes various relations into corresponding fact tables and dimension tables. The data model created using the multidimensional modeling is known as a data cube (Datta & Thomas, 1999). The data cube often uses a star schema. A star schema is composed of a single fact table and a dimension table for each dimension. The fact table contains foreign keys that connect it to other dimensions and quantitative measurements of a business activity under consideration. Dimension tables and their attributes are selected for their capability to contribute for further analysis.

2.3. Data mining technique

According to Frawley, Paitetsjy-Shapiro, and Matheus (1992), data mining is a process to discover uncertain, unknown, and hidden information from a database. Grupe and Owrang (1995) define data mining as a unique method of finding new facts and relationships in the existing data that have not as yet been discovered by experts. Berry and Linoff (1997) regard data mining as an analysis method using automatic or semi-automatic tools to discover the meaningful relationship or rules from a huge amount of data. However, three important features distinguish data mining technique from other statistics methods (Hand, 1998, 1999). First, it is able to categorize large amounts of data for secondary analysis. Second, it is able to abstract information from data from operating application systems, rather than based on prior experiments. Finally, it can discover patterns and relationships in data.

Fayyad (1996) distinguishes between knowledge discovery technique in database process and data mining technique. Knowledge discovery in database process is a series of processes on selecting appropriate data from the database. These processes include activities such as preprocessing function of missing data, transformation function of data, and evaluation function of discovered knowledge. Data mining, in contrast, is only one of the aforementioned steps among the entire process.

The aim of data mining is to discover hidden information from huge amounts of data to aid decision makers in making more intelligent and timely decisions (Wang, Tseng, & Liao, 2009). The most used methods in data mining are decision tree, neural network, inductive learning, Bayesian network, and association rule (Han & Fu, 1999). Furthermore, tax agencies in most countries have frequently adopted data mining tools to assist in identifying taxpayers who evade obligations (Micci-Barrera & Ramachandran, 2004). In Malaysia, the Inland Revenue Board also benefits from data mining tools. A self-assessment method with data mining technology for tax payments assists in efficiently generating revenue in their operations (Rahman, 2008).

2.3.1. Association rules data mining

The standard association rule mining technique discovers correlations among items of transactions by employing the Apriori algorithm for finding frequent item sets with available, preset support, and confidence.

An association rule can be used to find dependent attributes of records within the database (Agrawal, Imielinski, & Swami, 1993). The current study applies the association rule on VAT data to search for patterns or rules of relationships between the associated attributes of VAT evasion. The association rule is described as follows.

Let Database D be the collection of records. Each record contains certain attributes. Let $A = \{A_1, A_2, A_3, \ldots, A_m\}$ be a set of all attributes, where m is the total number of attributes in the records. In addition, T is a subset of A, denoted as $T \subseteq A$. An association rule thus has the form $A_1, A_2, \ldots, A_j \rightarrow A_k$, where $A_1, A_2, \ldots, A_j, A_k \subseteq T$. If it has enough records, a conclusion can be made that a relationship occurs under the condition of associative attributes. In learning association rules, support and confidence are two important parameters for judging the validity of an association rule. Support shows statistical significance of the rule, whereas confidence shows the strength of the rule. Support and confidence are defined as follows:

$$support(A_1, A_2, \cdots, A_j, A_k) \equiv P(A_1, A_2, \cdots, A_j, A_k) = \frac{\#(k - dependent\ \ Attributes)}{\#\{total\ \ Reports\}}$$

and

$$confidence(A_1, A_2, \cdots, A_j \rightarrow A_k) \equiv P(A_k | A_1, A_2, \cdots, A_j)$$

For a meaningful association rule, its support and confidence must be larger than acceptable thresholds (Kuo, Lin, & Shih, 2007). Support, or coverage, represents the percentage of records that satisfy the condition of associative attributes $A_1, A_2, \ldots, and\ A_j$, whereas confidence, or accuracy, represents a probability that A_k is satisfied under the condition of associative attributes$A_1, A_2, \ldots, and\ A_j$ is satisfied.

2.3.2. Mining association rules with categorical and numeric attributes

In discovering correlations among underlying tax records, an association rule can be viewed as being defined over attributes of a relation and has the form $A_i \wedge A_j \rightarrow A_k$, where A_i and A_j are conjunctions of conditions (Rastogi & Shim, 2002). In tax domains, most of the attributes are either categorical or numerical by nature. In its most generic form, an association rule has the form $A_i[l_{i1}, u_{i1}] \wedge A_j = v_j \rightarrow A_k = v_k$, where l_{i1} and u_{i1} are values of a range interval from partitioning domain of numeric attribute A_i., and v_j and v_k are either categorical or numeric values of attributes A_j and A_k, correspondingly.

If association rules are permitted to contain disjunctions in each condition, association rules can be further enriched to contain more complex conditions by using a combination of operators $\wedge$ or $\vee$ in a rule. This can be accomplished by allowing multiple disjunctions over a numeric attribute. For instance, in $(A_i[l_{i2}, u_{i2}]) \wedge A_j = v_j \rightarrow A_k = v_k$, the rule contains two disjunctions for the numeric attribute A_i. Using individual interval of a numeric attribute in an association rule may not have adequate minimum support or confidence. Aggregating the disjunction intervals of a numeric attribute increases the chance to meet the required support and confidence.

The size of a numeric attribute domain can be rather large. In order to reduce search space to perform association rule mining, a numeric attribute domain has to split into several non-overlapping intervals. Variant schemes are then used to partition the numeric domain, such as histogram construction (Jagadish et al., 1998) and clustering algorithm (Lent, Swami, & Widom, 1997).

3. System framework

The goal of the current study is to apply the association rules data mining technique to enhance the performance and/or productivity for VAT evasion detection in Taiwan. The reason for the selection of Taiwan data is that VAT is an important tax source in that country, ranking second only to income tax. Furthermore, VAT evasion is a serious issue in Taiwan (Huang & Lin, 2009).

The VAT reported data are originally stored in an Oracle operational database system. To avoid interfering with the tax system operation, data transform service (DTS) is employed to transform the data into a SQL server database. Data extraction, preprocessing, and transformation steps/activities are all executed in the SQL server database. The program Object Linking and Embedding Database (OLEDB) is used to process and load objects into different data marts. The SQL server is used to establish associated data cubes. DBMiner Enterprise Version is employed to implement the association rule data mining. The application ActiveX Data Objects (ADO) is also used for data access between data cubes and DBMiner. The entire system framework is illustrated in Fig. 1. Details of the framework are further discussed in the following subsections.

3.1. VAT data samples selection

The data samples were collected from the VAT database. Business entities identified as likely involved in VAT evasion activities were selected as data samples. The collected data samples were divided into two sets: training data and validation data. Association rule was applied on the training data set first to discover important patterns (i.e., rules of value-added tax evasion). The validation data set was used to validate the accuracy of the trained association rules or patterns. Another set of data samples excluded from training and validation data samples for testing was collected subsequently.

3.2. Data warehouse construction

In the proposed framework, different aspects of business entities were collected. The data warehouse was organized into three dimension tables: Tax Evasion Control File, Tax Registration File, and Tax Reports File. Each table was designed as a dimension table. The fact table and dimension tables formed a star schema in the data warehouse.

3.2.1. Attributes extraction

According to Hunter and Nelson (1996), the operation of tax evasion detection can be divided into two subfunctions: tax evasion prevention and tax evasion detection. The goal of tax prevention is to close the loophole of tax reporting. The goal of tax detection is to locate defective tax reports. Tax authorities commonly impose tax evasion punishments to business entities that illegally evade taxes. In screening tax reports for further auditing, absolute amounts such as sales amount per se may not be good criteria for selecting suspicious tax reports. Instead, the relative measures such as value-added ratio, defined as the ratio of costs divided by sales, are used to screen the tax reports suspected of illegal tax evasion practices.

The attributes extracted from different aspects of business entities are summarized in Tables 1–3. Most of the attribute domains are numeric data in either integers or real numbers. In order to improve the quality of the training model and reduce training time for conducting data mining procedure, the numeric data are transformed into categorical data. With the help of tax authorities, numeric attributes are partitioned into intervals, with each interval representing categorical data.

Table 1 contains information of business entities that have been identified as guilty of value-added tax evasion.

Table 2 reports the registration information of the business entities. Each business entity is classified into one of 38 business sectors by tax authorities upon its registration. When a business entity changes its business sector, it is required to register again.

Table 3 presents the business entity's bimonthly tax reports. Because this study does not directly use attribute values of VAT reports as the quantification indicators, some of the predefined ratios are calculated from original VAT reports and stored in the table as attributes. These ratio attributes are transformed into categorical data for improving the quality of training model.

3.2.2. Data marts and data cubes

Data marts are subsets of a data warehouse. Each data mart contains specific data on a specific subject. VAT data stored in a data warehouse are extracted into data marts according to the specific analytical requirements.

A data cube is used to organize data marts into a multidimensional database. A data cube consists of dimensions and facts. Dimensions represent perspectives of VAT subjects. Every dimension is associated with a dimension table. The dimension table contains the data of a specific subject. The fact table represents a center theme of multidimensional data model. The attributes violator-host-ban and business-entity-ban in the tables are foreign keys to relate the fact table. The key attributes provide the unique identification to link the fact table and dimensional tables in the data cube.

The association rule of DBMiner, an on-line analytical mining (OLAM) system, is employed as the data-mining tool in this study. The DBMiner software package uses multilevel association rules to implement association rule algorithm. The fast frequent-pattern tree algorithm (Han, Pei, Yin, & Mao, 2004) is used to process intra-dimensional association mining.

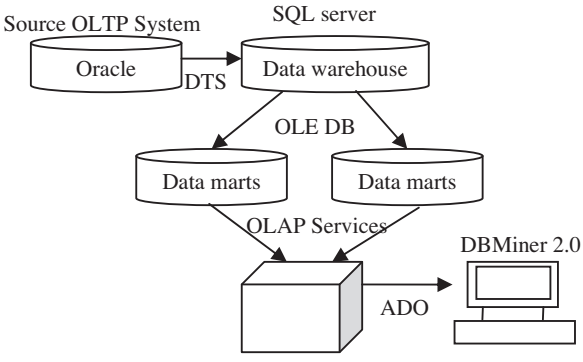

Fig. 1. System framework.

4. Screening model design

Visual Basic scripts were utilized to perform data sample selection and data preprocessing on SQL Server 7.0.

4.1. Data samples collection

The data samples were collected from the VAT database. The data of business entities suspected as VAT evasion cases were identified and selected for training and validating the screening model. Due to the sensitive nature of the study, only the data before 2006 were allowed for analysis. Data samples from 2003 and 2004 were used as training data set and validation data set, respectively. The number of data samples collected is listed in Table 4.

Table 1
Tax evasion control file.

	Attribute names	Definition
1	Violator-host-ban	Business entity's registration number
2	Amount of tax evasion	Identified Value-Added tax evasion amount in the tax report period. Categorized into 6 classes
3	Beginning date	Beginning date of a certain tax report period
4	Ending date	Ending date of a certain tax report period

Table 2
Tax registration file.

	Attribute names	Definition
1	Business-entity-ban	Business entity's registration number
2	Business category	Business entity's business sector
3	Capital	Total capital of the business entity

Table 3
Tax reports file.

	Attribute names	Definition
1	Business-entity-ban	Business entity's registration number
2	Year	Year of tax report
3	Month	Month of tax report
4	Sales amount	Total amount of sales in the certain tax report period Categorized into 6 classes
5	Sales-capital ratio	Sales amount in the certain tax report period divided by capital of the business entity Categorized into 14 classes
6	Value-added ratio	Difference of sales amount and costs divided by sales amount Categorized into 12 classes
7	Exemption ratio	Tax free sales amount divided by sales amount Categorized into 13 classes
8	Sales return or allowance ratio	Amount of sales return after sold or allowance divided by sales amount Categorized into 13 classes
9	Duplicate uniform invoice ratio	Duplicate uniform invoice divided by sales amount Categorized into 13 classes

Table 4
Data samples.

Year	Number of tax evasions	Number of tax reports	Number of tax registrations
2003	2130	3426	1934
2004	1650	2222	1543

Each sample represents a business entity's tax report that was identified as anomalous. Because VAT is reported bimonthly, six tax reports are filed per year for a business entity. Business entities are allowed to file another tax report if they find an error in a previous tax report within two weeks. As a result, a business entity having two tax reports in the same tax report period is possible. These duplicate tax reports are aggregated into a single tax report in the data-cleaning stage. During the auditing process, tax agents merge two or more tax reports into a single punishment case for the same business entity if these tax report periods are in sequence. This is the reason why the number of tax evasions is less than the number of tax reports, as shown in Table 4. The number of tax registrations is also less than the number of tax evasions because, for tax evasion purposes, some business entities have more than one tax report period identified.

The validation data set was further divided into three validation data subsets. Threefold cross-validation was applied to measure the average accuracy rate. Data samples of the first two months of 2005 were collected for testing purpose. Three hundred samples were available for testing.

4.2. Data cubes construction

Varieties of data cubes can be constructed from data warehouse for different analytical requirements. To study the dependency of attributes, Data Cube 1 and Data Cube 2 were constructed using different sets of attributes. In Data Cube 1, the attributes chosen from three dimensions included the amount of tax evasion in the Tax Evasion Control File, the business category and capital in the Business Tax Registration File, and the sales amount, valued-added ratio, and sales-capital ratio in the Business Tax Report File. The attributes of Data Cube 2 included the amount of tax evasion in the Tax Evasion Control File, the business category and capital in the Business Tax Registration File, and the exemption ratio, sales return or allowance ratio, and duplicate uniform invoice ratio in the Business Tax Report File. Attribute dependency of Data Cube 2 was expected to be stronger than that of Data Cube 1. Attribute dependency was verified by examining the accuracy rates of the screened validation samples.

4.3. Screening model

To develop the screening model, the association rule was utilized to mine tax evasion patterns from the training data set. First, the patterns consisted of a set of rules. Second, the validation data set was divided into three sets for threefold validation. Third, the rules were then validated by using the validation data sets. Validation was measured as the average accuracy rate representing the average of three different validation accuracy rates. After the average accurate rates of rules were calculated, rules with low accuracy rate were pruned. The remaining rules set was applied to screen tax reports suspected of tax evasion. The diagrams of training, validation, and testing of the proposed screening model are illustrated in Fig. 2.

5. Experimental results

Association rule method of DBMiner was utilized separately on Data Cube 1 and Data Cube 2 to obtain association rules. The number of VAT evasions was used as a parametric measurement for data mining results.

5.1. Support and confidence selection

Support and confidence were set for association rule mining in the range of 4–15% and 80–90%, respectively, with the number of association rules obtained from data mining results shown in Tables 5 and 6 for Data Cubes 1 and 2, respectively. The higher the support rate, the fewer rules discovered. Consequently, with a fixed support rate, increasing confidence rate results in fewer rules. When more rules were discovered, more tax reports were selected for auditing for possible tax evasion purpose. The tax audit staff can choose the suitable combination of support and confidence rates for screening according to their desired auditing plans. For illustration purpose, support and confidence of 7% and 85% for Data Cube 1, and 10% and 80% for Data Cube 2 were chosen to evaluate the accuracy rates.

5.2. Association rules generation

An association rule is a pattern of correlation between attributes in a taxpayer's report. Because training data samples are ta-

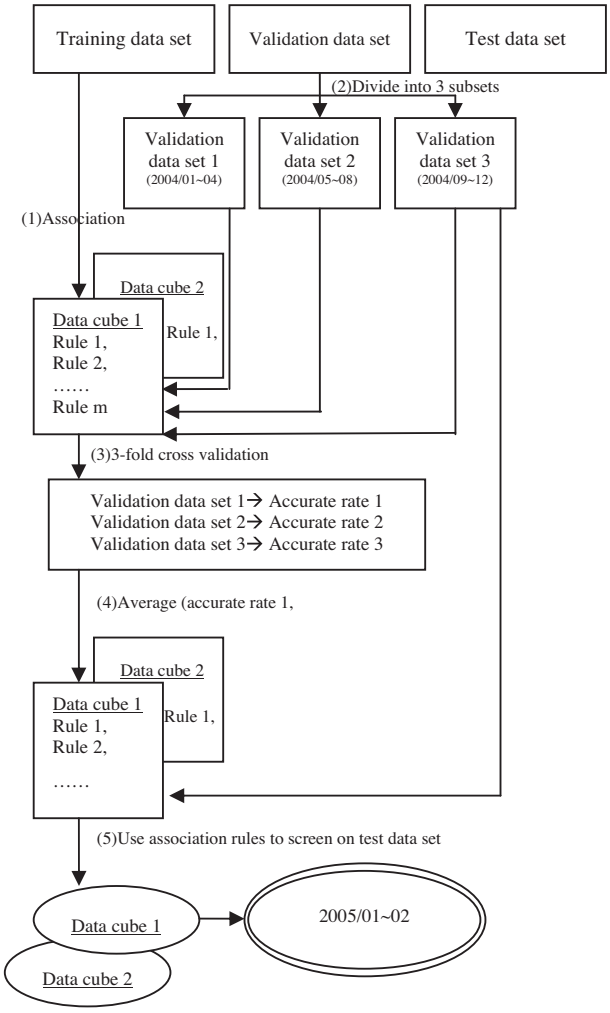

Fig. 2. Diagram of training, validation, and testing of the proposed screening model.

ken from business entities identified as involved with possible tax evasion, the association rules generated by data mining represent tax evasion patterns. For instance, the following association rule

$$BSCD38 = [03] AND sales\_cap\_class = [01] AND sales\_class = [01] \Rightarrow EVD\_TAX\_CLASS = [02]$$

is interpreted as "if a business entity's business category is buy-and-sell class (Class 03), sales capital ratio is in between 0 and 10 (Class 01), and sales amount is less than one million (Class 01), it is then identified as possible VAT evasion case and its amount of tax evasion is in between 0 and 100,000 dollars (Class 02)." The number in the bracket of the rule represents the attribute's categorical class.

Table 5
Data mining results on Data Cube 1.

Support (%)	Confidence		
	80% Number of rules	85% Number of rules	90% Number of rules
4	34	25	9
5	33	21	6
6	27	18	6
7	23	16	5
8	20	12	4
9	18	11	4
10	16	10	4
11	14	9	3
12	13	9	3
13	12	8	3
14	10	6	1
15	8	5	1

Table 6
Data mining results on Data Cube 2.

Support (%)	Confidence		
	80% Number of rules	85% Number of rules	90% Number of rules
4	14	2	4
5	11	1	4
6	10	0	4
7	10	0	4
8	10	0	4
9	10	0	2
10	10	0	0
11	8	0	0
12	8	0	0
13	8	0	0
14	8	0	0
15	7	0	0

The sample association rules generated by mining on Data Cubes 1 and 2 are listed in Tables 7 and 8, respectively. Originally, 16 rules were generated from Data Cube 1 and 10 rules were generated from Data Cube 2. After rules were posted and processed, subsumed and merged, eight rules were left for Data Cube 1 and five rules for Data Cube 2.

Two rules that have the same pattern can be subsumed into each other. For instance, the following two association rules represent the same pattern:

$$BSCD38 = [01] AND return\_sales = [00] AND xmp\_sales = [00] \Rightarrow EVD\_TAX\_CLASS = [01]$$

and

$$BSCD38 = [01] AND return\_sales = [00] AND EVD\_TAX\_CLASS = [01] \Rightarrow xmp\_sales = [00].$$

Two association rules with adjacent attribute classes are merged into a single rule. For instance,

$$sales\_cap\_class = [01] AND valued\_rate = [08] \Rightarrow EVD\_TAX\_CLASS = [01]$$

and

$$sales\_cap\_class = [01] AND valued\_rate = [09] \Rightarrow EVD\_TAX\_CLASS = [1]$$

are merged into an association rule as follows:

$$sales\_cap\_class = [01] AND valued\_rate = [08 \sim 09] \Rightarrow EVD\_TAX\_CLASS = [1]$$

Merging adjacent attribute classes is equivalent to combining multiple disjunction intervals of the attribute. That two individual rules may not meet minimum support or confidence in screening the related tax reports is highly possible. Aggregating intervals of an attribute will thus increase the chance for the association rule to have minimum required support or confidence.

5.3. Association rules validation

In order to measure the effectiveness of derived association rules, the association rules were applied onto three validation data sets. Accuracy rates were calculated separately for each validation

Table 7
Association rules generated from Data Cube 1 with support 7% and confidence 85%.

Body	Implies	Head	Support (%)	Confidence (%)
BSCD38=[01] AND sales_cap_class=[01] AND sales_class=[01]	==>	EVD_TAX_CLASS=[01]	13.46	91.34
sales_cap_class=[01] AND sales_class=[01] AND valued_rate =[11]	==>	EVD_TAX_CLASS=[01]	8.89	88.46
BSCD38=[01] AND sales_class=[01] AND valued_rate=[07]	==>	EVD_TAX_CLASS=[01]	16.94	91.44
sales_cap_class=[01] AND sales_class=[01∼02]	==>	EVD_TAX_CLASS=[01]	8.55	89.84
sales_cap_class=[01] AND valued_rate[08∼09]	==>	EVD_TAX_CLASS=[01]	9.13	87.09
Sales_class=[01] AND valued_rate [12]	==>	sales_cap_class=[01]	10.75	89.10
sales_class=[01∼02]	==>	EVD_TAX_CLASS=[01]	36.31	88.09
valued_rate[07∼10]	==>	EVD_TAX_CLASS=[01]	15.43	85.81

Table 8
Association rules generated from Data Cube 2 with support 10% and confidence 80%.

Body	Implies	Head	Support (%)	Confidence (%)
BSCD38=[01∼02] AND c2cv_sales=[00] AND EVD_TAX_CLASS=[01]	==>	xmp sales=[00]	10.13	80.62
BSCD38=[01] AND return_sales=[00] AND xmp sales=[00]	==>	EVD_TAX_CLASS=[01]	27.42	84.51
c2cv_sales=[00] AND return_sales=[00] AND xmp sales=[00]	==>	EVD_TAX_CLASS=[01]	22.16	80.14
BSCD38=[02] AND c2cv_sales=[00] AND	==>	EVD_TAX_CLASS=[01]	10.33	80.91
c2cv_sales=[00] AND EVD_TAX_CLASS=[01]	==>	return_sales=[00]	22.51	80.39

Table 9
Validation result of Data Cube 1 (average accurate rate 87.69%).

	Validation Data Set 1 (2004/01–04)		Validation Data Set 2 (2004/05–08)		Validation Data Set 3 (2004/09–12)	
	Manual screening (A)	Data mining (B)	Manual screening (A)	Data mining (B)	Manual screening (A)	Data mining (B)
Rule 1		95		116		96
Rule 2		39		36		23
Rule 3		128		117		116
Rule 4		103		130		130
Rule 5		34		26		35
Rule 6		20		17		16
Rule 7		39		35		28
Rule 8		55		53		60
Total	586	513	602	530	576	504
Accuracy rate (B/A) × 100%	87.54%		88.04%		87.50%	

data set. The validation accuracy rate is taken from the average of the three individual accuracy rates. The validation results of Data Cubes 1 and 2 are shown in Tables 9 and 10, respectively.

From the technical point of view, applying association rules onto each validation data set enables the counting of how many manual screening data samples are selected by the association rules. The current screening method used by tax authorities is manual screening. The manual screening counts were provided by the tax authority. The numbers used in the data mining columns are the counts selected by each association rule. Duplicate data samples were eliminated if the data samples were already selected by other rules. The accuracy rate is calculated as a ratio of the total selected data samples to manual screening samples. For instance, the Validation Data Set 1 for Data Cube1 has 586 samples, with 513 samples meeting at least one of eight association rules, and the accuracy rate is 87.54%. This study also takes average of different accuracy rates on multiple validation data sets to smoothen the outlier data samples. During validation process, association rules that either selected too many duplicate samples or made little contribution in screening process were discarded.

Comparing the accurate rates from either each validation data set or average, the accuracy rates of Data Cube 2 are higher than those of Data Cube 1. In the proposed experiment design of data cubes, the dependency of the attributes in Data Cube 2 is assumed to be stronger. These results confirm that the selection of high dependency attributes for the data cube will result in producing better association rules.

5.4. Test on association rules

To calculate the accuracy rate on the testing data set, derived association rules were applied to the testing data set of VAT reports filed during the first two months of 2005. The accuracy rates were estimated by counting the tax reports identified as involved in tax

Table 10
Validation result of Data Cube 2 (average accurate rate 95.97%).

	Validation Data Set 1 (2004/01–04)		Validation Data Set 2 (2004/05–08)		Validation Data Set 3 (2004/09–12)	
	Manual screening (A)	Data mining (B)	Manual screening (A)	Data mining (B)	Manual screening (A)	Data mining (B)
Rule 1		76		55		80
Rule 2		199		207		193
Rule 3		48		44		55
Rule 4		97		101		99
Rule 5		138		174		127
Total	586	558	602	581	576	554
Accuracy rate (B/A) × 100%	95.22%		96.51%		96.18%	

Table 11
Testing result using 2005/01–02 data samples.

	Test data: number of tax reports	No. of tax evasions		
		Manual screening (A)	Data mining screening (B)	Hit ratio (B/A) × 100%
Data cube 1	166,005	1017	746	73.35%
Data cube 2	166,005	1017	856	84.17%

Table 12
Error rate of Data Cube 1 using 2005/01–02 data samples.

Manual	Data mining screening Prediction cases		Total
	Number of tax evasions	Number of non-tax evasions	
Actual cases			
Number of tax evasions	394	22	416
Number of non-tax evasions	352	249	601
Total	746	271	1017

Table 13
Error rate of Data Cube 2 using 2005/01–02 data samples.

Manual	Data mining screening Prediction cases		Total
	Number of tax evasions	Number of non-tax evasions	
Actual cases			
Number of tax evasions	401	15	416
Number of non-tax evasions	445	146	601
Total	856	161	1017

evasion by manual screening process and selected by the proposed data mining screening model. The accuracy rates, as shown in Table 11, are 73.35% and 84.17% for Data Cubes 1 and 2, respectively.

The error rate of the manual screening method and prediction of the proposed data mining method were further analyzed. The error predictions are summarized in Tables 12 and 13.

For Data Cubes 1 and 2, the risk rates of no selection of tax evasion using data mining screening are 5.29% (22/416 × 100%) and 3.61% (15/416 × 100%), respectively. These risk rates represent tax losses. Compared with the manual screening method that only selects from portion of tax reports, the proposed data mining screening method has lesser tax-loss risk. For non-tax evasion selected by the proposed data mining screening method, the risk rates are 58.57% (352/601 × 100%) and 73.88% (445/601 × 100%), respectively, which is not significantly different from 59.09% (601/1017 × 100%). These risk rates can be interpreted as opportunistic costs. The opportunistic costs represent the associated staffing expenses and other resources spent on auditing the compliant tax reports.

6. Conclusions

The goal of the current study is to use data mining techniques to identify and select suspicious VAT evasion reports for further auditing. Compared with the manual screening method, the proposed data mining technique is a more scientific and resource-saving approach. Using the data mining technique on a large amount of tax data to derive tax evasion patterns can improve the accuracy rates in screening potential tax evasion reports. Thus, the data mining method can be employed to screen all tax reports and reduce the unnecessary wasting of auditing staff resources.

The data mining screening model is designed to help tax auditors personnel perform their tax evasion screening tasks more efficiently, thereby enhancing the productivity of auditing possible tax evasion cases. In addition, by informing taxpayers that their tax reports can be quickly and scientifically analyzed, voluntary compliance rates are expected to improve. This will save valuable resources and improve recovery of tax losses.

The current study has three contributions. First, the data mining tool can be supported for filtering possible non-compliant VAT reports. Instead of relying on manual methods and personal judgments in selecting suspicious tax reports, tax authorities now have a more scientific way of identifying possible evaders. Second, although limited studies utilize mining association rules to detect tax evasion, the mining outcome with association rules presented in the current study provides a direction for future research within tax field. Third, the current study has identified specific patterns and significant features of illegal taxpayers. Thus, the tax auditors can combine this method with their professional experience to detect further cases of tax evasion.

However, the current study has some limitations. Due to budget limitation, the current study used IBM DBMiner 2.0 as the data mining tool, rather than more advanced software. Other data mining software might be able to identify a more effective association rule to improve tax evasion detection performance. Moreover, this study only filtered out suspicious tax evasion case without processing real auditing.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD conference on management of data* (pp. 207–216). Washington, DC, USA.

Berry, M. J. A., & Linoff, G. (1997). *Data mining technique: For marketing sale, and customer support*. Hoboken, NJ: John Wiley & Sons.

Datta, A., & Thomas, H. (1999). The cube data model: A conceptual model and algebra for on-line analytical processing in data warehouses. *Decision Support Systems, 27*, 289–301.

Fadairo, S. A., Williams, R., Trotman, R., & Onyekelu-Eze, A. (2008). Using data mining to ensure payment integrity. *Journal of Government Financial Management, 57*, 22–24.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*, 20–25.

Fayyad, U. M., Piatesky-Shapiro, G., & Padhraic, S. (1996). From data mining to knowledge discovery: An overview. In *Advance in knowledge discovery and data mining* (pp. 1–34). American Association for Artificial Intelligence, Menlo Park, CA, USA.

Frawley, W. J., Paitetsjy-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine, 13*, 57–70.

Gebauer, A., Nam, C. W., & Parsche, R. (2007). Can reform models of value added taxation stop the VAT evasion and revenue shortfalls in the EU? *Journal of Economic Policy Reform, 10*, 1–13.

Grupe, F. H., & Owrang, M. M. (1995). Data base mining discovering new knowledge and cooperative advantage. *Information Systems Management, 12*, 26–31.

Gupta, R. (2008). Tax evasion and financial repression. *Journal of Economics and Business, 60*, 517–535.

Han, J., & Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions of Knowledge and Data Engineering, 11*, 798.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery, 8*, 53–87.

Hand, D. (1998). Data mining: Statistics and more? *The American Statistician, 52*, 112–118.

Hand, D. (1999). Statistics and data mining: Intersecting disciplines. *ACM SIGKDD Explorations, 1*, 16–19.

Hoover, J. N. (2009). States use BI, data warehousing to recoup unpaid taxes. *Intelligent Enterprise*, 12.

Huang, M., & Lin, S. (2009). Tax evasion and audit selection. *International Journal of Accounting Studies, 48*, 35–66.

Hunter, W. J., & Nelson, M. A. (1996). An IRS production function. *National Tax Journal, 49*, 105–115.

Inmon, W. H. (1996). *Building the data warehouse*. New York: John Wiley & Sons, Hoboken.

Jagadish, H., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K. & Suel, T. (1998). Optimal histograms with quality guarantees. In *Proceedings of the 24th VLDB conference*, New York.

Keen, M., & Lockwood, B. (2010). The value added tax: Its causes and consequences. *Journal of Development Economics, 92*, 138–151.

Kuo, R. J., Lin, S. Y., & Shih, C. W. (2007). Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. *Expert System with Application, 33*, 794–808.

Lent, B., Swami, A., & Widom, J. (1997). Clustering association rules. Technical Report. Stanford InfoLab. Thirteenth International Conference on Data Engineering, April 7–11, Birmingham, UK.

Liao, S. H. (2003). Knowledge management technologies and applications- literature review from 1995 to 2002. *Expert System with Application, 25*, 155–164.

Micci-Barrera, D., & Ramachandran, S. (2004). Improving tax administration with data mining. Available at: www.spss.com.

Ministry of Finance. (2011). Monthly statistics of finance. Available at: http://www.mof.gov.tw/engweb/ct.asp?xItem=53723&CtNode=683&mp=2.

Nycz, M., & Smok, B. (2006). Data warehouse – The source of business information. In *Proceedings of the 2006 information science and IT education Joint Conference* (pp. 243–250). Salfork, UK.

Rahman, M. S. A. (2008). Utilisation of Data mining technology within the accounting information system in the public sector: a country study – Malaysia. Doctoral Dissertation.

Rastogi, R., & Shim, K. (2002). Mining optimized association rules with categorical and numeric attributes. *IEEE Transaction of Knowledge and Data Engineering, 14*, 29–50.

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*, 491–500.

Songini, M. L. (2004). Fraud sniffers. *Computerworld*, 38.

Wang, Y. H., Tseng, M. H., & Liao, H. C. (2009). Data mining for adaptive learning sequence in English language instruction. *Expert System with Application, 36*, 7681–7686.

Yoon, Y. (1999). Discovery knowledge in corporate databases. *Information System Management, 16*, 64–71.