Expert Systems with Applications 40 (2013) 1427–1436

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Characterization and detection of taxpayers with false invoices using data mining techniques

Pamela Castellón González [a], Juan D. Velásquez [b,*]

[a] Servicio de Impuestos Internos, Government of Chile, Chile
[b] Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box 8370439, Chile

ARTICLE INFO

Keywords:
False invoices
Fraud detection
Data mining
Clustering
Prediction

ABSTRACT

In this paper we give evidence that it is possible to characterize and detect those potential users of false invoices in a given year, depending on the information in their tax payment, their historical performance and characteristics, using different types of data mining techniques. First, clustering algorithms like SOM and neural gas are used to identify groups of similar behaviour in the universe of taxpayers. Then decision trees, neural networks and Bayesian networks are used to identify those variables that are related to conduct of fraud and/or no fraud, detect patterns of associated behaviour and establishing to what extent cases of fraud and/or no fraud can be detected with the available information. This will help identify patterns of fraud and generate knowledge that can be used in the audit work performed by the Tax Administration of Chile (in Spanish *Servicio de Impuestos Internos* (**SII**)) to detect this type of tax crime.

1. Introduction

Tax evasion and tax fraud[1] have been a constant concern for tax administrations, especially when pertaining to developing countries (Davia, Coggins, Wideman, & Kastantin, 2000). While it is true that taxes are not the only source of government funding, the fact is that they send a very important signal about the commitment and effectiveness with which the State can carry out its functions and restrict access to other sources of income.

In particular, the value added tax (VAT), implemented in over 130 countries at different stages of economic development has become a key component of tax revenues, raising about 25% of the world's tax revenue (Harrison & Krelove, 2005). In the case of Chile, taxes provide about 75% of the resources from which the State each year pays its expenses and investments, collecting during 2011 a total of USD $41.6 billion dollars.[2] VAT represents 45% amounting to USD $18.7 billion dollars and generating over 400 million invoices a year, of which 56% is issued in paper format and 44% in electronic format (Bergman, 2010).

The phenomenon of false invoices in respect of VAT is explained by the mechanics of determining the tax payable. When a company receives a false invoice, it simulates a purchase that never existed, thus increasing its tax credit fraudulently and decreasing VAT payment. Also, there is a decrease of payment in the income tax due to increased costs and expenditures declared.

The falsity of the document may be *material* if the physical elements that make up the invoice have been adulterated, or *ideological* when the materiality of the document is not altered, but the operations recorded in it are adulterated or nonexistent. The latter is more complex and difficult to detect because it involves fictitious transactions in which an audit is required to examine the sales books and corrections, or cross referencing the information with suppliers. Moreover, these cases are more expensive for SII, as they require a greater amount of time dedicated to collecting and testing evidence, which is harder to find.

The best known cases of material falsification are the physical adulteration of the document, the use of *hanging* invoices in which an invoice is counterfeited to impersonate a taxpayer of good behavior, and the use of a double set of tax invoices, which has two same-numbered invoices, but one of which is fictional and for a higher amount. In ideological falsification, invoices are used to register a nonexistent operation or adulterate the contents of an existing operation.

According to a method used by the SII to estimate VAT evasion (Schneider & Enste, 2000) resulting from false invoices and other credit enlargements applied in the period 1996–2004, evasion by false invoices has historically represented between 15% and 25% of total VAT evasion, increasing significantly in years of economic

* Corresponding author. Tel.: +56 2978 4834; fax: +56 2689 7895.
E-mail address: jvelasqu@dii.uchile.cl (J.D. Velásquez).
URL: http://wi.dii.uchile.cl/ (J.D. Velásquez).

[1] Usually refers to *tax avoidance* when referring to behaviors that, within the law, prevent or reduce taxes, while *evasion or tax fraud* involves a violation of the law to obtain the same results.

[2] Considering only Central Government tax revenue (excluding CODELCO, the main government cooper miner company, municipalities and social security).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.051

crisis. This is why in the crisis of 1998–1999 the participation rate increased to 38%, reaching an amount close to USD $1 billion dollars. This becomes relevant since recently there was a global economic crisis that hit Chile in late 2008 and the middle of 2009, causing an increase in the rate of VAT evasion to 23%, in the amount of evasion of USD $4 billion dollars.

It also requires that resources be invested in well-focused monitoring, detecting those taxpayers who have greater compliance risk and not bother or waste time and resources on those who do comply (Slemrod & Yitzhaki, 2002). For this, data mining techniques offer great potential, because they allow the extraction and generation of knowledge from large volumes of data to detect and characterize fraudulent behavior and failure to pay tax, in the end improving the use of resources (Fayyad, Piatestky-Shapiro, & Smyth, 1996).

This paper is organized as follows. Section 2 describes how artificial intelligence techniques have facilitated the detection of tax evasion in tax administrations. Section 3 describes the data mining techniques applied. Section 4 describes the type of information used and the main results obtained in the characterization and detection of fraud in the issuance of invoices, and Section 5 presents the main conclusions and future lines of research.

2. Related work

Fraud in its various manifestations is a phenomenon that no modern society is free of. All governments, regardless of whether they are large or small, public or private, local or multinational, are affected by this reality, which seriously undermines the principles of solidarity and equality of citizens before the law and threatens business.

There are many fields and industries affected by this phenomenon. A study conducted by Chena, Huang, and Kuo (2009) in 2006, surveyed 150 medium and large Chilean companies to consult on this issue. The results show that 41% of them were victims of fraud in the past two years. This poses great challenges in prevention and opportunities for detection (Bonchi, Giannotti, Mainetto, & Pedreschi, 1999), given that fraud is usually higher than reported by companies, because somehow disturbs the image of the company towards customers and suppliers. In many cases there are even companies that are not known to have been victims of fraud.

Many fraud detection problems involve a large amount of information (Lundin, Kvarnstomr, & Jonsson, 2003). Processing these data in search of fraudulent transactions requires a statistical analysis which needs fast and efficient algorithms, among which data mining provides relevant techniques, facilitating data interpretation and helping to improve understanding of the processes behind the data (Myatt Glenn, 2007). These techniques have facilitated the detection of tax evasion and irregular behavior in other areas such as banking, telecommunications, insurance, IT, money laundering, and in the medical and scientific fields, among others (Cechhini, Aytug, Koehler, & Pathak, 2010).

To detect tax fraud, tax institutions began using random selection audits or focusing on those taxpayers who had no previous audits in recent periods and selecting cases according to the experience and knowledge of the auditors. Later methodologies were developed based on statistical analysis and construction of financial or tax ratios which evolved into the creation of rule-based systems and risk models (OECD, 1999). These transform tax information into indicators which permit ranking of taxpayers by compliance risk. In recent years, the techniques of data mining and artificial intelligence have been incorporated into the audit planning activities (US Government Accountability Office, 2004; OECD, 2004b), mainly to detect patterns of fraud or evasion, which are used by tax authorities for specific purposes.

The internal revenue service (**IRS**), the institution responsible for administering taxes in the United States, has used data mining techniques (US Government Accountability Office, 2004) for various purposes, among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities (Dubin, 2007), electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering (OECD, 2004a; OECD, 2004b; Watkinsa et al., 2003). The Fig. 1 shows part of the data mining techniques used by tax administrations as well as the logistic regression models, decision trees, neural networks, clustering algorithms and visualization techniques such as link analysis.

In the Australian Tax Office, the *Compliance Program* is based on a risk model which uses statistical techniques and data mining in order to make comparisons, to find associations and patterns by logistic regression, decision trees and SVM (US Government Accountability Office, 2004; US Government Accountability Office, 2008). A case of interest has been the approach used by Denny and Christen (2007), of discovering small clusters or unusual subpopulations, called *Hot Spots*, using techniques such as the self organizing map (SOM) to explore its features, clustering algorithms like k-means and visuals that are easy to understand for non-technical users.

In New Zealand, the existing model associates the degree of compliance with attention to auditing, which coincides with that used by the Australian counterpart (OECD, 2004a). The plan includes an analysis of the economic, international, population, ethnic diversity and family structure. For its part, Canada uses neural networks and decision trees to distinguish the characteristics of taxpayers who evade or commit fraud, based on the results of past audits, to detect patterns of noncompliance or evasion (OECD, 2004b).

In Latin America, Peru was one of the first to apply these techniques to detecting tax evasion (García & Valderrama, 2007; Torgler, 2005), adding to the selection system of the Maritime Customs of Callao an artificial intelligence tool based on neural networks. During 2004, this model was improved through the application of fuzzy rules and association for pre-processing variables and classification and regression trees (CART) to select the most relevant variables.

Brazil has developed project risk analysis and applied artificial intelligence (HARPIA) jointly with the Brazilian Federal Revenue and universities in the country (Digiampietri et al., 2008). This project consists of developing a detection system of atypical points to help the regulators to identify suspicious transactions based on a graphic display of information on historical imports and exports and a system of export product information based on Markov chains, to help importers in the registration and classification of their products, avoid duplication and to calculate the probability that a string is valid in a given domain.

In the case of Chile the first trial was developed in 2007 (Lückeheide, Velásquez, & Cerda, 2007), using the SOM and k-means to segment VAT taxpayers according to their F29 statements and characteristics. Later, in 2009, following the international trend, risk models were built of different stages of the life cycle of the taxpayer, in which neural networks, decision trees and logistic regression techniques are applied. The first trial was further developed to identify potential users of false invoices through artificial neural networks and decision trees, mainly using information from tax and income declarations in micro and small enterprises.

3. Data mining techniques applied

For purposes of characterization and identification of patterns three data mining techniques are applied: self organizing maps