International Journal of Accounting Information Systems 15 (2014) 357–380

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits☆

Glen L. Gray [a,*], Roger S. Debreceny [b]

[a] David Nazarian College of Business and Economics, California State University at Northridge, United States
[b] School of Accountancy, Shidler College of Business, University of Hawai'i at Mānoa, United States

ARTICLE INFO

Article history:
Received 1 June 2013
Received in revised form 26 December 2013
Accepted 15 May 2014
Available online 5 July 2014

Keywords:
Auditing
Fraud
Data mining

ABSTRACT

This paper explores the application of data mining techniques to fraud detection in the audit of financial statements and proposes a taxonomy to support and guide future research. Currently, the application of data mining to auditing is at an early stage of development and researchers take a scatter-shot approach, investigating patterns in financial statement disclosures, text in annual reports and MD&As, and the nature of journal entries without appropriate guidance being drawn from lessons in known fraud patterns. To develop structure to research in data mining, we create a taxonomy that combines research on patterns of observed fraud schemes with an appreciation of areas that benefit from productive application of data mining. We encapsulate traditional views of data mining that operates primarily on quantitative data, such as financial statement and journal entry data. In addition, we draw on other forms of data mining, notably text and email mining.

1. Introduction

This study explores the targeted application of data mining techniques to fraud detection as a core component of financial statement audits.[1] Data mining refers to the extraction of knowledge from large

☆ The authors thank Sev Grabski and Eckhardt Kriel for their valuable comments and suggestions, as discussants of an earlier version of this paper presented at the 2013 University of Waterloo Research Symposium on Information Integrity & Information Systems Assurance.

* Corresponding author.
E-mail addresses: glen.gray@csun.edu (G.L. Gray), roger@debreceny.com (R.S. Debreceny).

[1] While the paper focuses on the detection of fraud within the financial statement audit conducted by external auditors, most of the key messages are also relevant for internal auditors.

http://dx.doi.org/10.1016/j.accinf.2014.05.006
1467-0895/

volumes of data (Han and Kamber, 2006. 5). Data mining involves acquisition, loading and integration of data; application of specialist data mining tools and, finally, human interpretation of the discovered meaning.[2] The decision to incorporate data mining into financial audits is both a firm-level decision for accounting firms and an engagement-level decision. Firm-level decisions preclude engagement-level decisions in that if firm management does not see a beneficial reason to invest resources in software, infrastructure, training, and staffing then data mining will likely not be a cost-effective option for engagement teams. Larger accounting firms and some specialist providers offer a variety of data mining services. Currently, data mining is used in specialized audits (e.g., fraud audits or forensic audits) by expert staff in the professional services firms; however, data mining is seldom used in financial statement audits. When used, it is for identified high-risk clients by the firm's data mining specialists. The majority of this paper is focused on engagement-level data mining activities; however, we revisit firm-level issues and research opportunities in the concluding parts of the paper.

Applying data mining to fraud detection as part of a routine financial audit can be challenging and, as we will explain, data mining should be used when the potential payoff is high. In general, when it comes to fraud detection for a given audit client, the audit team would make three major decisions: (1) What specific types of fraud (e.g., revenue recognition, understated liabilities, etc.) should be included in the audit plan for a particular client? (2) What sources of data (e.g., journal entries, emails, etc.) would provide evidence of each type of fraud? (3) Which data mining technique(s) (e.g., directed or undirected techniques) would be the most effective for finding potential evidence of fraud in the selected data? Developing answers for each of these questions is significant individually, but, in combination, answering these questions is challenging. These challenges may encourage the audit team to continue to use traditional – but less diagnostic – analytical and substantive procedures. However, as we will discuss in this paper, each of the populations for each of these three questions can be intelligently reduced so that the application of data mining to fraud detection becomes more manageable and will have a higher potential for a successful payoff. We also recognize that data mining techniques and associated software can have a steep learning curve. Further, if used improperly, data mining can produce many false positives and spurious patterns that will require auditors to expend time to subsequently investigate. The primary contribution of this paper is in identifying specific fraud and evidence combinations where data mining would be the most effective in traditional financial audits as well as those combinations where data mining would be least effective. Identifying the more effective use of data mining could encourage auditors to include data mining as a regular element of their audit programs. Future researchers can build on our exploratory findings to further refine the application of data mining in financial statement audits.

Specifically, the paper proposes a taxonomy that includes three components, namely, account schemes and evidence schemes (components of fraud schemes as defined by Gao and Srivastava (2011)),[3] and data mining functionality to identify the most effective combinations of those three components. Data mining has the potential to enhance the efficiency and effectiveness of the audit. Productive data mining toolsets are now more widely available and auditors have access to a cornucopia of audit-relevant data both internal and external to the client organization. Internal data can include financial data, non-financial data, and email archives. Externally, a vast array of quantitative and qualitative information on organizations is now available on the Internet and in commercial financial and textual databases. These include news reports, blog postings, Facebook postings, and Twitter feeds. For public companies, regulatory filings such as the filings made on the U.S. Securities and Exchange Commission's (SEC's) EDGAR database in XBRL format are available.

There has been an increased interest in data mining for fraud detection in the regulatory and professional domain. For example, the SEC has developed an "Accounting Quality Model," (commonly referred to as the Robocop) designed to identify anomalous financial statement filings to the Commission. The tool mines the XBRL data repository along with other datasets (Lewis, 2012; Rohman and Berg, 2013). A recent report by the Financial Executives International (FEI) also points to a variety of tools to data mine the XBRL filings to the SEC (FERF, 2013). The Advisory Committee on the Auditing Profession (ACAP) to the U. S. Treasury recommended

[2] CRISP-DM sets out a methodology that sets out a process model that sees data mining flowing from the development of a business context for the mining; understanding data sources; data preparation; modeling; evaluation and deployment (Chapman et al., 2000).

[3] Gao and Srivastava (2011) divided fraud schemes into two components: account schemes reflecting the accounts impacted by the fraud (e.g., fictitious revenue) and evidence schemes reflecting how the fraudster implemented the fraud (e.g., fake documents).