ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

A Survey of Data Mining-based Fraud Detection

R.Anitha

Asst. Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore, India

ABSTRACT: This survey paper categories, compares, and summaries from different articles in automated fraud detection. It defines the professional fraudster, focuses the main types and subtypes of known fraud, and presents the nature of data evidence collected within affected industries. Within the business context of mining the data to achieve higher cost savings, this research presents methods and techniques together with their problems. Compared to all reviews on fraud detection, this survey covers more technical articles and is the only one, to the best of our knowledge, which proposes alternative data and solutions from related domains.

KEYWORDS: Data mining applications, automated fraud detection, adversarial detection.

I. INTRODUCTION

Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data. This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model.

The term fraud here refers to the abuse of a profit organization's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe.

Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.

There are often two main criticisms of data mining-based fraud detection research: the dearth of publicly available real data to perform experiments on; and the lack of published well-researched methods and techniques. To counter both of them, this paper garners all related literature for categorisation and comparison, selects some innovative methods and techniques for discussion; and points toward other data sources as possible alternatives.

Section 2 – Who are the white collar criminals which a fraud detection system should be designed to discover? Where can one apply data mining techniques to commercial fraud? Section 3 – What analytical methods and techniques from other adversarial domains can one apply in fraud detection? Section 4 – How is this fraud detection survey different from others? Section 7 concludes with a brief summary

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

II. RELATED WORK

This section highlights the types of fraudsters.

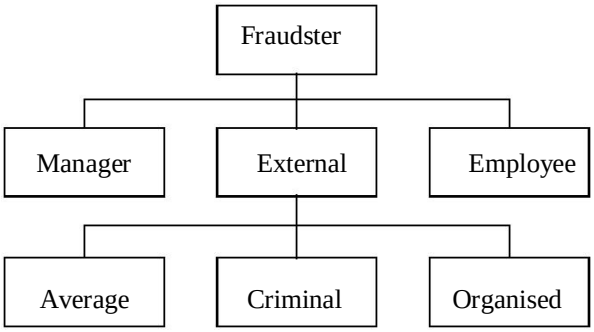

Figure 2.1: Hierarchy chart of white-collar crime perpetrators from both firm-level and community-level perspectives.

From figure 2, the fraudster can be an external party, or parties. Also, the fraudster can either commit fraud in the form of a prospective/existing customer (consumer) or a prospective/existing supplier (provider). The external fraudster has three basic profiles: the average offender, criminal offender, and organised crime offender. Average offenders display random and/or occasional dishonest behaviour when there is opportunity, sudden temptation, or when suffering from financial hardship.

With reference to figure 2, the profit-motivated fraudster has interactions with the affected business. Traditionally, each business is always susceptible to internal fraud or corruption from its management (high-level) and non-management employees (low-level). In addition to internal and external audits for fraud control, data mining can also be utilised as an analytical tool.

In contrast, the more risky external fraudsters are individual criminal offenders and organised/group crime offenders (professional/career fraudsters) because they repeatedly disguise their true identities and/or evolve their modus operandi over time to approximate legal forms and to counter detection systems. Therefore, it is important to account for the strategic interaction, or moves and countermoves, between a fraud detection system's algorithms and the professional fraudsters' modus operandi. It is probable that internal and insurance fraud is more likely to be committed by average offenders; credit and telecommunications fraud is more vulnerable to professional fraudsters.

The main purpose of these detection systems is to identify general trends of suspicious/fraudulent applications and transactions. In the case of application fraud, these fraudsters apply for insurance entitlements using falsified information, and apply for credit and telecommunications products/services using non-existent identity information or someone else's identity information. In the case of transactional fraud, these fraudsters take over or add to the usage of an existing legitimate credit or telecommunications account.

III. METHODS AND TECHNIQUES

This section examines four major methods commonly used, and their corresponding techniques and algorithms. Volume of both fraud and legal classes will fluctuate independently of each other; therefore class distributions (proportion of illegitimate examples to legitimate examples) will change over time. Multiple styles of fraud can happen

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

at around the same time. Each style can have a regular, occasional, seasonal, or once-off temporal characteristic. Legal characteristics/behaviour can change over time. Within the near future after uncovering the current modus operandi of professional fraudsters, these same fraudsters will continually supply new or modified styles of fraud until the detection systems start generating false negatives again.

the common data mining approaches to determine the most suspicious examples from the incoming data stream (evaluation data) are:

1) Labelled training data (A + B + C + D) can be processed by single supervised algorithm. A better suggestion is to employ hybrids such as multiple supervised algorithms or both supervised and unsupervised algorithms to output suspicion scores, rules and/or visual anomalies on evaluation data.

2) All known legal claims/applications/transactions/accounts/ sequences (C) should be used processed by semi-supervised algorithms to detect significant anomalies from consistent normal behaviour.

However, there are many criticisms with using labelled data to detect fraud:

> In an operational event-driven environment, the efficiency of processing is critical. The length of time needed to flag examples as fraudulent will be the same amount of time the new fraud types will go unnoticed. The class labels of the training data can be incorrect and subject to sample selectivity bias. They can be quite expensive and difficult to obtain. Staffs have to manually label each example and this has the potential of breaching privacy particularly if the data contains identity and personal information.

Therefore it is necessary to:

3) Combine training data (the class labels are not required here) with evaluation data (A + C + E + F). These should be processed by single or multiple unsupervised algorithms to output suspicion scores, rules and/or visual anomalies on evaluation data

IV. OTHER ADVERSARIAL DOMAINS

This section explains the relationship between fraud detection three other similar domains.

Terrorist Detection

There had been simplistic technical critiques of data mining for terrorist detection such as low accuracy (unacceptably high false positive rates in skewed data) and serious privacy violations (massive information requirements). To counter them, Jensen *et al* (2003) recommend fixed-size clustering to generate true class labels and the linked structure of data. Scores are randomly drawn from either the negative or positive entities' normal distributions. The second-round classifier averages an entity's first-round score and scores of all its neighbours. To reduce false positives, results show that second-round classifier reduces false positive rates while maintaining true positive rates of first-round classifier. To reduce information requirements, results show moderately high accuracy through the use of only twenty percent of the data.

Surveillance systems for terrorist, bio-terrorist, and chemo-terrorist detection often depend on spatial and spatio-temporal data. These are unsupervised techniques highly applicable to fraud detection. Neill and Moore (2004) employ Kulldorff's spatial scan statistic and the overlap-kd tree data structure. It efficiently finds the most significant densities from latitude and longitude of patient's home in real emergency department, and zip codes in retail cough and cold

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

medication sales data. Das *et al* (2004) utilise Partially Observable Markov Decision Process (POMDP) with Kulldorff's spatial scan statistic on to detect artificial attacks from real emergency department's spatio-temporal data.

Financial Crime Detection

Financial crime here refers to money laundering, violative trading, and insider trading and the following are brief application descriptions which correspond to each type of monitoring system for the United States government. The Financial Crimes Enforcement Network AI System (FAIS) (Senator *et al*, 1995) operates with an expert system with Bayesian inference engine to output suspicion scores and with link analysis to visually examine selected subjects or accounts. Supervised techniques such as case-based reasoning, nearest neighbour retrieval, and decision trees were seldom used due to propositional approaches, lack of clearly labelled positive examples, and scalability issues. Unsupervised techniques were avoided due to difficulties in deriving appropriate.

Intrusion and Spam Detection

There are multiple data sources for intrusion detection and the common ones are at host level, network level, and user level. Otey *et al* (2003) advocates intrusion detection at the secure Network Interface Cards (NIC) level. Leckie and Yasinsac (2004) argue that in an encrypted environment, intrusion detection can utilise metadata at the user level. Julisch and Dacier (2002) mine historical alarms to reduce false positives. Shavlik and Shavlik (2004) monitor Windows® operating systems. Compared to fraud detection where many studies use real proprietary data, the benchmark KDD cup 1999 network intrusion detection data is often used.

V. CONCLUSION

This survey has explored almost all published fraud detection studies. It defines the adversary, the types and subtypes of fraud, the technical nature of data, performance metrics, and the methods and techniques. After identifying the limitations in methods and techniques of fraud detection, this paper shows that this field can benefit from other related fields. Specifically, unsupervised approaches from counterterrorism work, actual monitoring systems and text mining from law enforcement, and semi-supervised and game-theoretic approaches from intrusion and spam detection communities can contribute to future fraud detection research. However, Fawcett and Provost (1999) show that there are no guarantees when they successfully applied their fraud detection method to news story monitoring but unsuccessfully to intrusion detection. Future work will be in the form of credit application fraud detection

REFERENCES

1. Artis, M., Ayuso M. & Guillen M. (1999). Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market. *Insurance Mathematics and Economics* **24**: 67-81.
2. Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). Synthesizing Test Data for Fraud Detection Systems. *Proc. of the 19th Annual Computer Security Applications Conference*, 384-395.
3. Belhadji, E., Dionne, G. & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance* **25**(4): 517-538.
4. Bell, T. & Carcello, J. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. *Auditing: A Journal of Practice and Theory* **10**(1): 271-309.
5. Beneish, M. (1997). Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance. *Journal of Accounting and Public Policy***16**: 271-309.
6. Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. *Proc. of GECCO2000*.
7. Bentley, P., Kim, J., Jung., G. & Choi, J. (2000). Fuzzy Darwinian Detection of Credit Card Fraud. *Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society*.
8. Bhargava, B., Zhong, Y., & Lu, Y. (2003). Fraud Formalisation and Detection. *Proc. of DaWaK2003*, 330-339.
9. Bolton, R. & Hand, D. (2002). Statistical Fraud Detection: A Review (With Discussion). *Statistical Science* **17**(3): 235-255.
10. Bolton, R. & Hand, D. (2001). Unsupervised Profiling Methods for Fraud Detection. *Credit Scoring and Credit Control VII*.