Decision Support Systems 53 (2012) 543–553

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

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

An experimental comparison of real and artificial deception using a deception generation model

Yanjuan Yang [a], Michael V. Mannino [b,*]

[a] Automapath Inc., Santa Clara, CA 95050, United States
[b] The Business School, University of Colorado Denver, United States

ARTICLE INFO

Article history:
Received 6 April 2011
Received in revised form 23 March 2012
Accepted 29 April 2012
Available online 5 May 2012

Keywords:
Deception
Deception detection
Noise model
Data generation model

ABSTRACT

To develop a data mining approach for a deception application, data collection costs can be prohibitive because both deceptive data and truthful data are necessary to be collected. To reduce data collection costs, artificially generated deception data can be used, but the impact of using artificially generated deception data is not well understood. To study the relationship between artificial and real deception, this paper presents an experimental comparison using a novel deception generation model. The deception and truth data were collected from financial aid applications, a document centric area with limited resources for verification. The data collection provided a unique data set containing truth, natural deception, and boosted deception. To simulate deception, the Application Deception Model was developed to generate artificial deception in different deception scenarios. To study differences between artificial and real deception, an experiment was performed using deception level and data generation method as factors and directed distance and outlier score as outcome variables. Our results provided evidence of a reasonable similarity between artificial and real deception, suggesting the possibility of using artificially generated deception to reduce the costs associated with obtaining training data.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Deception involving intentionally provided false information can lead to adverse outcomes in many areas including law enforcement, national security, employment, government benefits, taxation, and university admissions. Because of the impact of deception on decision making, there has been an increasing interest in learning about deception and its detection for many years. Research has mainly focused on detecting deception in richly mediated communication channels with more recent emphasis on deception in text and documents.

This research involves document deception, an emerging area of concern in industry and government. In document deception, an individual falsifies an application for an obligation such as a tax liability or a benefit such as a position, financial aid, loan, government benefit, or admission to a university. The most prominent area of document deception, tax fraud by individuals and corporations, has been a long standing concern of government. The U.S. I.R.S. estimates the net tax gap of $385B in 2006 [17], an increase of $85B from the estimate of the 2001 tax gap. With the ease of submitting electronic applications, fraud in other areas such as health care reimbursements, mortgage applications, and welfare applications have grown in importance in recent years. According to CoreLogic Inc., loan origination fraud was estimated at $12B in 2010 and $7.4B in 2011 [10]. The decline in loan origination fraud is due to reduced loan origin amounts and tighter lending standards in the mortgage industry enacted as a result of growing mortgage application fraud in the mid-2000s.

With the increasing cost of higher education, financial aid deception has become a prominent form of document deception. Federal, state and private financial aid programs target assistance toward students with the least ability to pay for college. This targeting of aid is based on self-reports of financial condition by students and parents. Honest reporting of financial condition in student financial aid applications ensures equitable allocation of scarce financial aid resources. Colleges and universities routinely verify the accuracy of a subset of aid applications. According to the report prepared by Rhodes and Tuccillo [26], 30% of dependent student records and 20% of independent student records had false data fields when schools verified the information as part of the random sample process. The results of discrepancies in the original financial aid applications were estimated to cause improper payment of approximately $270 million (15.9%) of U.S. Pell dollars in 2006–2007.

This research is primarily motivated by the unavailability of data and cost of data collection for developing data mining methods to

* Corresponding author at: Campus Box 165, P.O. Box 173364, University of Colorado Denver, Denver, CO 80217-3364, United States.
E-mail addresses: juanmaoy@yahoo.com (Y. Yang), Michael.Mannino@ucdenver.edu (M.V. Mannino).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.04.009

detect document deception. According to [24], unavailability of data sets is a major deterrent to developing data mining approaches for fraud detection. In most deception studies, deceptive data paired with truth data are collected manually [6,35]. In a university setting, internal review boards may restrict data collection in studies involving deception increasing the cost and difficulty to obtain a suitable data set. In the student financial aid area, financial aid offices perform tedious audits of applications to determine discrepancies between truth and non-compliant data. To lower the cost of data collection, artificially generated deception data can be used to train a data mining program, but the impact of using artificially generated deception data is not well understood. For new systems in development, the availability and reliability of test data with deception may be severely limited. Generating artificial data may be the only available way to test a new system on deceptive cases.

There is limited understanding about the relationship between artificially generated deception and real deception. A number of studies [1,19,22,23,38] have used artificially generated noise to study the sensitivity of classification algorithm performance to noise with little understanding of the relationship between real and artificial noise. In the intrusion detection domain, large amounts of artificial test data were generated for the 1998 and 1999 US DARPA competition [14] although no evaluation of the generated data was reported. For fraud detection in video on demand usage, Barse et al. [2] developed a data generation methodology and performed evaluation. However, their results were limited by the authentic data collection, deception model, and informal comparison between authentic and synthetic data. Thus, previous research does not provide a reasonable understanding about the relationship between real deception and artificially generated deception in document-centric deception.

The goals of this study are to develop a deception generation model and to investigate the fit between real deception data and artificial deception data created with the deception generation model. To simulate deception, the Application Deception Model (ADM) was developed to generate artificial deception in different deception scenarios. The ADM substantially extends previous data generation approaches through goal directed changes for groups of related attributes in observations with incentive to deceive. Deception data and the ground truth data were collected from financial aid applications, a document-centric area with limited resources for verification. The data collection provided naturally occurring deception in which subjects had some incentive to falsify applications and boosted deception in which subjects were instructed to falsify their applications. Using the collected data and the ADM, an experiment was conducted with deception level and data generation method as factors and directed distance and outlier score as outcome variables. The experimental results indicated a reasonable fit between artificially generated deception and real deception, suggesting the possibility of using artificially generated deception to train data mining algorithms.

This paper makes three contributions to research and practice. First, this paper emphasizes the difference between noise and deception through development of a deception generation model and empirical comparison between artificially generated noise and deception. Most previous research has failed to make this distinction. Second, this paper brings rigor to the study of artificially generated deception through a careful empirical comparison of the data characteristics. Previous research has not performed careful empirical comparisons of artificially generated deception. Third, the results of this study suggest the value of artificially generated deception in practice. More research is needed to confirm the value and understand limitations in specific applications. If the value of artificially generated deception is confirmed, lower data collection costs may allow improved deception detection policies and methods to be developed.

The rest of the paper is organized as follows. In the next section, we briefly review deception and noise background and provide details about past efforts to generate artificial deception. Section 3 presents the data collection process for the real deceptive data set used in the study. Section 4 describes the data generation models used in the study particularly the Application Deception Model (ADM), a novel method of artificial deception generation developed for this study. Section 5 presents the experiment design to investigate the relationship between real deception and artificial deception generated by data generation models. The experimental results and findings are presented in Section 6. Section 7 concludes the study.

2. Related work

The theoretical foundation for this research is drawn from a combination of theories of deception and noise. To provide a context for this study, some literature about deception and its detection are briefly reviewed. The directly relevant efforts on the usage of artificially generated noise and deception are presented after the deception background.

Deception detection has a long history especially in adversarial areas such as law enforcement and intelligence gathering. Numerous studies have noted that the accuracy with which people typically identify deception is only slightly better than chance (approximately 54%) [4]. This issue is more intense when the deception is conveyed in documents because of the lack of nonverbal cues.

Practice in detection of document deception is dominated by scoring models to target audit resources. Because these models are not public, red flags and guidelines have emerged to guide individuals seeking to avoid audit. For example, FraudGuides.com provides a list of likely triggers for an IRS audit. A relatively small number of classification methods have been proposed in the literature for detection of document deception. Bonchi et al. [3] proposed a classification-based methodology for constructing profiles of fraudulent taxpayers in tax fraud detection. More recent work by Thang et al. [30] has used fuzzy inference and neural network for tax fraud detection in small businesses. There has also been an initial attempt at applying decision trees in fraud detection for border customs processing [27]. Extensive research has also been conducted on methods of health care fraud detection [21].

Although deception and noise both involve deviations from the true state of an attribute, the underlying data generation processes are different. This research effort focuses on deception as intentional misrepresentations and noise as random, unintentional deviations. Although some research exists about unintentional deception [29], document-based deception involves effort to complete a form with specific field values. Prominent studies about noise in a data mining context focus on unbiased random noise [13,22,38]. To provide contrast, we compared a standard noise model used in data mining studies to the deception generation model developed for this research.

In the literature, some research has treated deception as noise. Jiang et al. [18] conducted a study to handle explicitly noisy input data on the web. Although the authors refer to noise, the study actually deals with deception on the web. To cope with deception, they proposed two methods: knowledge base modification (KM) and input modification (IM). The KM method modifies the knowledge base (a decision tree) to account for distortion in the inputs provided by the user. The IM method modifies an observed input to the most likely true value of the input given the observations made by the system. They used a distortion rate parameter to generate artificial training and testing data. The distortion rate parameter does not support goal directed state changes for deception. In addition, the work assumes that inputs are distorted independently rather than allowing scenarios in which groups of inputs are jointly modified.

The most closely related research [2] involves synthetic data generation and evaluation for fraud detection in video on demand

Download English Version:

https://daneshyari.com/en/article/554759

Download Persian Version:

https://daneshyari.com/article/554759

Daneshyari.com