Decision Support Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

# Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud

Eugen Stripling[a,*], Bart Baesens[a,b], Barak Chizi[c], Seppe vanden Broucke[a]

[a] Department of Decision Sciences and Information Management, KU Leuven, Leuven, Belgium
[b] School of Management, University of Southampton, Southampton, UK
[c] Department of Information Systems and Software Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

ARTICLE INFO

*Keywords:*
Workers' compensation insurance fraud
Fraud detection
Conditional anomaly detection
Isolation forest

ABSTRACT

The development of new data analytical methods remains a crucial factor in the combat against insurance fraud. Methods rooted in the research field of anomaly detection are considered as promising candidates for this purpose. Commonly, a fraud data set contains both numeric and nominal attributes, where, due to the ease of expressiveness, the latter often encodes valuable expert knowledge. For this reason, an anomaly detection method should be able to handle a mixture of different data types, returning an anomaly score meaningful in the context of the business application.

We propose the *i*Forest$_{CAD}$ approach that computes conditional anomaly scores, useful for fraud detection. More specifically, anomaly detection is performed conditionally on well-defined data partitions that are created on the basis of selected numeric attributes and distinct combinations of values of selected nominal attributes. In this way, the resulting anomaly scores are computed with respect to a reference group of interest, thus representing a meaningful score for domain experts. Given that anomaly detection is performed conditionally, this approach allows detecting anomalies that would otherwise remain undiscovered in unconditional anomaly detection.

Moreover, we present a case study in which we demonstrate the usefulness of our proposed approach on real-world workers' compensation claims received from a large European insurance organization. As a result, the *i*Forest$_{CAD}$ approach is greatly accepted by domain experts for its effective detection of fraudulent claims.

## 1. Introduction

Across all lines of insurance, it is conservatively estimated that fraud causes a monetary damage of $80 billion a year [1]. Given this estimate, it is self-evident that insurance fraud is a major problem that adversely affects our society [2]. Among the various insurance lines, the Insurance Information Institute [3] (or short III), reported that the majority of industry experts (69%) believes in an increase of workers' compensation (WC) insurance fraud. WC is an insurance policy to cover costs that emerge when employees sustain an injury or become ill on the job. Fraudsters view the deprivation of money from insurance organization as a *low-risk*, *high-reward game*, since it is far safer than other money earning, serious crimes such as armed robbery or drug trafficking [1, 4]. It should therefore be of no surprise that even when considering WC insurance alone, the total loss caused by WC fraud can easily reach tens of millions of dollars [5]. Taking these points into consideration, it is therefore strongly advised by the III [6] to invest in the advances of analytical technology to protect the insurance organization and their honest clients against the ever-changing nature of increasingly intricate fraud practices.

In this paper, we propose a novel analytical approach, called *i*Forest$_{CAD}$, that performs isolation-based anomaly detection *conditionally* on reference groups (i.e., data partitions) meaningful to domain experts. The resulting *i*Forest$_{CAD}$ anomaly scores are then leveraged for fraud detection. Data partitions are defined by distinct combinations of values of selected nominal attributes, thereby integrating a mixture of nominal and numeric attributes in a meaningful way. Based on the observation that fraud data sets usually consist of both nominal and numeric attributes [7], our proposed *i*Forest$_{CAD}$ approach aims to fulfill the strong desire to make use of all available information in the combat against fraud.

Moreover, we present a case study in which we apply our *i*Forest$_{CAD}$ approach on a data set of real-world WC claims received from a large European insurance organization. For the study, we collaborated with

* Corresponding author at: Department of Decision Sciences and Information Management, Naamsestraat 69 - box 3555, Leuven 3000, Belgium.
*E-mail addresses:* eugen.stripling@kuleuven.be (E. Stripling), bart.baesens@kuleuven.be (B. Baesens), chizba@bgu.ac.il (B. Chizi), seppe.vandenbroucke@kuleuven.be (S. vanden Broucke).

https://doi.org/10.1016/j.dss.2018.04.001
Received 12 November 2017; Received in revised form 17 April 2018; Accepted 17 April 2018

the insurer's special investigation unit (SIU) to fruitfully incorporate valuable expert knowledge of the private investigators (PIs) to enhance the automatic detection of fraudulent WC claims. One of the most important project goals in order to reach acceptance among the PIs is that the fraud detection approach satisfies the following major requirement:

Interestingness: It is to be ensured that suspected fraud cases reported to the PIs are *interesting*. This implies that the reported claims should conform with the PIs' expert knowledge (i.e, detection of fraudulent claims in line with known fraud patterns). Yet, the reported claims should also have some element of novelty and surprise (i.e., discovery of previously unseen fraud patterns).

The implications for the data scientist of the main requirement can be stated as follows:

1. **Integration of expert knowledge**: The inclusion of accumulated expert knowledge into the fraud detection mechanisms plays an essential role in order for the data science application to be accepted by the PIs.
2. **Accuracy**: Due to scarce resources, an efficient deployment of PIs to check and invest suspicious claims is required. Hence, the fraud detection model should be accurate in its predictions so that the PIs first focus their attention to the truly fraudulent claims.
3. **Explainability**: The data scientist *must* be able to explain to the PIs *why* the classification model predicts a claim as fraudulent.
4. **Novelty**: To reach acceptance among the PIs, the fraud detection approach should return fraudulent claims according to known patterns, but also detect novel ones that comply with the expertise of the PIs.

Since the PIs ultimately decide whether or not an in-depth investigation has to be conducted, it is crucial that the fraud detection approach fulfills the aforementioned criteria. In particular, special attention needs to be dedicated to the first criterion, because it is often beneficial to inject expert knowledge into the data analytical approach (see, e.g., [8-10]).

According to the PIs, given the type of injury and other information, an “unusual long” recovery time (or, equivalently, disproportional duration of incapacity), is a strong indicator of a WC claim being fraudulent. To capture this insight in a data-driven manner, one needs to answer the following questions: *How to decide when a recovery time is too long (without requiring human judgment)? How can valuable expert knowledge be integrated into the decision model construction?*

With an interesting real-world case study on WC fraud, we present a fraud detection approach (Fig. 1) that allows detecting claims with a disproportional recovery time in a *fully data-driven* manner by which information of mixed type attributes is processed in a way meaningful to the PIs.

We thereby leverage the well-established anomaly detection algorithm called isolation forest (*i*Forest), introduced by [11, 12]. The application of anomaly detection plays a crucial role as it allows for the *automatic* detection of disproportional recovery times. The *i*Forest algorithm is a key component of our proposed approach which we favor over other anomaly detectors for reasons we elaborate on in Section 2.2. It is important to note that the anomaly scores of the observations are computed conditionally on data partitions, which are defined based on the distinct combination of values of selected nominal attributes. Thus, the name of our proposed approach, $i\text{Forest}_{\text{CAD}}$, roots in the fact that a conditional anomaly detection (CAD) is performed with the aid of the *i*Forest. The created $i\text{Forest}_{\text{CAD}}$ scores are combined with the remaining attributes which then serve as an input for training a supervised classification model. In this way, we exploit the benefits of both supervised and unsupervised learning. The $i\text{Forest}_{\text{CAD}}$ scores proved to be indeed of high importance for the detection of suspicious claims in our case study.

Our research contributions can be summarized as follows:

- We propose the $i\text{Forest}_{\text{CAD}}$ method that computes anomaly scores conditionally on given reference groups (i.e., well-defined data partitions). In this way, $i\text{Forest}_{\text{CAD}}$ is capable of identifying “hidden” anomalies, which we demonstrate in Section 3.4.
- Our approach allows processing a mixture of nominal and numeric attributes, returning a condensed score that is meaningful in the context of the business application. The scores produced by $i\text{Forest}_{\text{CAD}}$ can be used not only for conditional anomaly detection but also as a new numeric input attribute for a predictive model.
- We demonstrate the application of anomaly detection and predictive analytics in the scope of a real-world case study on WC insurance fraud. To the best of our knowledge, the practical application of primarily machine learning techniques to combat WC fraud has not yet been presented in the literature.

The remainder of this paper is structured as follows. The next section provides more background information on WC insurance fraud, discusses anomaly detection and methods, as well as explains the inner workings of the *i*Forest algorithm. Section 3 formally introduces the concepts of our proposed approach with particular focus on the creation of $i\text{Forest}_{\text{CAD}}$ scores. Additionally, in Section 3.4, we provide an example that showcases the detection of hidden anomalies. Section 4 presents the case study in which we demonstrate the usefulness of the $i\text{Forest}_{\text{CAD}}$ scores for the detecting of fraudulent WC claims. In the same section, we elaborate on the effectiveness of applying conditional anomaly detection and how our proposed approach is applied to meet the most important requirement of interestingness. Section 5 summarizes the main findings of our work and highlights potential research directions.

## 2. Preliminaries

### 2.1. Workers' compensation fraud

Workers' compensation (WC) insurance provides a cost coverage in case employees sustain a work-related injury or disease that occur as a result of performing their occupational duties [3, 13]. For example, in the USA, coverage may be required for costs such as wage replacement, medical care and rehabilitation, and death benefits for the dependents if the employee deceased in work-related accidents (including terrorist attacks) [3].

According to the latest issue update on insurance fraud [6], it is believed that WC is one of the most vulnerable insurance lines to fraud. Further, III reported that 69% of industry experts forecast an increase in WC fraud. This strongly suggests initiating appropriate measures in order to protect insurance organizations and their honest clients against fraudsters. To do so, III pointed out that advances in analytical technology are a crucial factor in order to be able to keep up with the ever-changing nature of increasingly complex and sophisticated fraud schemes.

Viaene and Dedene [4] characterized insurance fraud by the presence of (at least) the following elements: (1) Misrepresentation of circumstances or material facts in the form of lie, falsification, or concealment, (2) deliberate plan of deception, and (3) purpose to gain unauthorized benefits. The authors further classified insurance fraud into three broad categories: (1) internal versus external, (2) underwriting versus claim, and (3) soft versus hard.

The first category (*internal versus external*) attempts to distinguish between the various types of perpetrators. Internal fraud is committed from within the insurance organization, e.g., by insurers, agents, and insurer employee. External fraud is perpetrated by individuals outside the organization, e.g., by applicants, policyholders, and claimants. The distinction sometimes becomes blurry in situations that involve a