Applied Soft Computing 36 (2015) 283–299

Contents lists available at ScienceDirect

Applied Soft Computing

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

An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance

Ilker Kose [a,*], Mehmet Gokturk [a,1], Kemal Kilic [b,2]

[a] Department of Computer Engineering, Gebze Technical University, 41400 Kocaeli, Turkey
[b] Faculty of Engineering and Natural Sciences, Sabanci University, 34956 Istanbul, Turkey

ARTICLE INFO

Article history:
Received 22 June 2013
Received in revised form 26 April 2015
Accepted 5 July 2015
Available online 30 July 2015

Keywords:
Decision support systems
Fraud detection
Health insurance
Interactive machine learning
Machine intelligence

ABSTRACT

Detecting fraudulent and abusive cases in healthcare is one of the most challenging problems for data mining studies. However, most of the existing studies have a shortage of real data for analysis and focus on a very limited version of the problem by covering only a specific actor, healthcare service, or disease. The purpose of this study is to implement and evaluate a novel framework to detect fraudulent and abusive cases independently from the actors and commodities involved in the claims and an extensible structure to introduce new fraud and abuse types. Interactive machine learning that allows incorporating expert knowledge in an unsupervised setting is utilized to detect fraud and abusive cases in healthcare. In order to increase the accuracy of the framework, several well-known methods are utilized, such as the pairwise comparison method of analytic hierarchical processing (AHP) for weighting the actors and attributes, expectation maximization (EM) for clustering similar actors, two-stage data warehousing for proactive risk calculations, visualization tools for effective analyzing, and *z*-score and standardization in order to calculate the risks. The experts are involved in all phases of the study and produce six different abnormal behavior types using storyboards. The proposed framework is evaluated with real-life data for six different abnormal behavior types for prescriptions by covering all relevant actors and commodities. The Area Under the Curve (AUC) values are presented for each experiment. Moreover, a cost-saving model is also presented. The developed framework, i.e., the eFAD suite, is actor- and commodity-independent, configurable (i.e., easily adaptable in the dynamic environment of fraud and abusive behaviors), and effectively handles the *fragmented* nature of abnormal behaviors. The proposed framework combines both proactive and retrospective analysis with an enhanced visualization tool that significantly reduces the time requirements for the fact-finding process after the eFAD detects risky claims. This system is utilized by a company to produce monthly reports that include abnormal behaviors to be evaluated by the insurance company.

1. Introduction

In traditional machine learning applications, the domain experts (who are usually the *end users* as well) usually participate in the modeling stage at two points [1,2]. First, the domain expert educates the knowledge engineers who develop the decision support tool so that the tacit knowledge (e.g., the objectives of the machine learning application, the relevant features that should be utilized in the analysis, natural groupings that are already known by the experts, and how to address missing data, etc.) can also be included in the development process. Secondly, especially for applications where the readily available data lack the required labels for machine-learning-based classification algorithms (i.e., unsupervised learning), the domain experts are asked to label the training data. Both of these contributions are commonly used and have proven to be very successful in various real-life applications.

However, in certain domains such as defense, healthcare, biosciences, etc., the experts are highly motivated and want to interact with the data and tools; hence, they are not satisfied by the above-mentioned role [1,3]. Therefore, they are usually reluctant to use traditional machine learning systems in their analysis unless they engage in the development process and are able to customize it

* Corresponding author. Tel.: +90 216 681 53 05; fax: +90 212 531 75 55; mobile: +90 532 645 24 34.
E-mail addresses: ikose@medipol.edu.tr (I. Kose), gokturk@gyte.edu.tr (M. Gokturk), kkilic@sabanciuniv.edu (K. Kilic).
[1] Tel.: +90 262 605 22 09; mobile: +90 533 425 83 95.
[2] Tel: +90 216 483 95 96; mobile: +90 505 211 82 99.

http://dx.doi.org/10.1016/j.asoc.2015.07.018
1568-4946/

based on their expertise. Furthermore, although computers are able to analyze vast amounts of data, human intelligence might be still preferable for analyzing smaller data sets in more detail [1,4,5].

As a result, the concept of interactive machine learning (IML) has emerged and attracted the attention of the research community and practices due to its ability to incorporate the domain experts directly into the model building process by providing various human–computer interaction tools and data visualization and analysis techniques as part of the developed applications [6–8]. Using IML, users can train a learning algorithm by choosing parameters, then evaluate and compare models, selecting those that are most appropriate for their goals.

Unlike the traditional machine learning approach, in which the humans (users) and machine work independently on different tasks, in IML, they work on the same task during the training stage [1,5]. That is, the human component utilizes the computational power of the machines to learn from the relationships hidden in data (using data visualization and analysis techniques) and in return directs the training process. This interaction makes IML particularly valuable whenever the hypotheses and objectives of machine learning are subject to change. The users can interact with the execution of the tool and modify it within the deployed environment.

This current research is motivated by the demand for a reliable and usable tool for fraud detection in the healthcare insurance business. In this study, an IML-based decision support tool that couples claim management systems, which will be utilized to detect fraud and abuse, is developed. The decision support tool uses the transactional data in order to identify suspicious cases (by assigning a value-function-based risk measure to each transaction) and provides a visual environment to aid users in the determination of whether the transaction is actual fraud or abuse.

Although there are some studies on anomaly detection time-series data and event sequence data [9] note that it is not viable to decide whether a *single transaction* is fraudulent or abusive in healthcare claims other than the trivial cases (such as prescribing a postpartum drug to a male patient, etc.) [10]. Usually, fraud and abuse in healthcare claims can only be detected if the earlier transactions by the same *actors* are also taken into consideration during the analysis. On top of the fruitless process of trying to label the transactions as fraudulent or abusive, the abundant number of transactions also limits the applicability of more elaborate classical machine learning techniques as the engine of the developed decision support tool. Furthermore, the fraudulent and abusive behavior evolves over time [9,11]. That is to say, the *actors* of the system are intelligent and adapt to the policing of insurance claims by changing their tactics. Therefore, rather than classical machine learning techniques, an IML-based decision support tool is extremely well suited for detecting fraud and abuse in healthcare insurance.

The developed framework imitates the process that experts use to determine suspicious cases, which is usually based on a *bottom-up* analysis of each actor and his or her relationship with healthcare associated commodities. Next, a *top-down* approach is utilized to automate the experts' method to identify the relevant evidence. The developed decision support tool significantly decreases the necessity of manual analysis by highlighting only the most suspicious cases and eliminating those that are unlikely to be critical. Furthermore, the data visualization tool enables the user to investigate each case effectively and thereby learn more about fraud and abusive behaviors, informing modifications of the risk assessment and evaluation engine.

The rest of the study is organized as follows. In Section 2, we present the problem statement and the ecosystem of the fraud and abuse detection problem. This section is critical for readers who are not familiar with the healthcare insurance domain. Furthermore, it will also clarify why a single, perfectly normal-seeming transaction might in fact be part of a greater fraud. Section 3 is devoted to the review of the related literature. Note that the existing literature, which is subtle but growing, deviates from our research in that it focuses only on the fraudulent and abusive behavior of a particular *actor* and not the transactions, which incorporate multiple actors and commodities. The developed framework is introduced in detail in Section 4. In order to measure the performance of the framework, an experimental analysis based on real data is presented in Section 5. The study ends with some concluding remarks and future research plans.

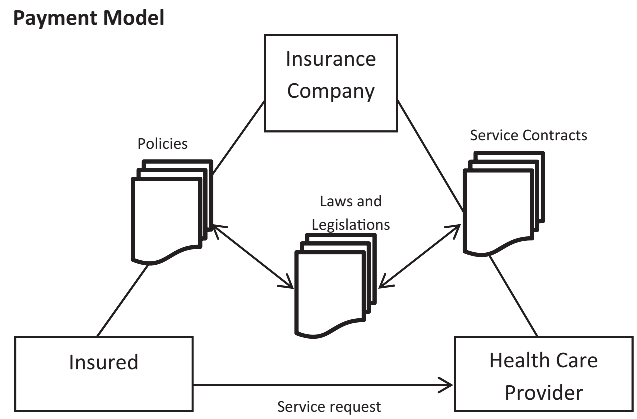

Fig. 1. The health insurance payment model.

2. The problem definition

The steady increase in life expectancy due to advancements in the health sciences, better standards of living, and increasing awareness of healthy lifestyles significantly enlarged the magnitude and share of healthcare in the global economy. For example, the Center for Medicare and Medicaid Services (CMS) reported that the total healthcare spending for the United States of America in 2010 was 2.6 trillion USD, which is 17.9% of the nation's gross domestic product (GDP). For comparison, in 2000, the US's total healthcare spending was 1.4 trillion USD (nearly half of its spending in 2010), and its share of the nation's GDP was 13.8% [12]. Improvements in medical technologies have also lead to major process and product diversification in healthcare services. Increase in healthcare expenditures, diversified healthcare services, and the rising expectations of the insured have cultivated the need for a systematic approach to manage the entire process. As a result, *healthcare insurance payment management systems* (for public and private insurance) have been developed to maintain the competitiveness of insurance companies and organizations in the marketplace.

Any healthcare insurance payment management system includes *the payment model* (e.g., fees for services, fees for capita) and *the claim management system*. The specifications regarding the payment, which are set by exogenous decision-makers, i.e., regulatory bodies, such as the government and the insurance authorities, primarily determine *the payment model*. Fig. 1 depicts the payment model which governs the payment among the three major stakeholders of the healthcare insurance systems, namely the insured, the insurance company and the healthcare providers. In addition, the *claim management system* component monitors the policy and contract terms (which are represented by the *payment rules*) among the stakeholders in order to provide proper healthcare service to the insured at the prices specified in the contract.

Generally, in practice, the management of the payment rules is considered as a relatively easy task and handled mostly by a