Expert Systems with Applications 38 (2011) 1039–1052

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An expert system for detecting automobile insurance fraud using social network analysis

Lovro Šubelj *, Štefan Furlan, Marko Bajec

Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1001 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Fraud detection
Automobile insurance
Social network analysis
Link analysis
Assessment propagation

ABSTRACT

The article proposes an expert system for detection, and subsequent investigation, of groups of collaborating automobile insurance fraudsters. The system is described and examined in great detail, several technical difficulties in detecting fraud are also considered, for it to be applicable in practice. Opposed to many other approaches, the system uses networks for representation of data. Networks are the most natural representation of such a relational domain, allowing formulation and analysis of complex relations between entities. Fraudulent entities are found by employing a novel assessment algorithm, *Iterative Assessment Algorithm* (*IAA*), also presented in the article. Besides intrinsic attributes of entities, the algorithm explores also the relations between entities. The prototype was evaluated and rigorously analyzed on real world data. Results show that automobile insurance fraud can be efficiently detected with the proposed system and that appropriate data representation is vital.

1. Introduction

Fraud is encountered in a variety of domains. It comes in all different shapes and sizes, from traditional fraud, e.g. (simple) tax cheating, to more sophisticated, where entire *groups* of individuals are collaborating in order to commit fraud. Such groups can be found in the automobile insurance domain.

Here fraudsters stage traffic accidents and issue fake insurance claims to gain (unjustified) funds from their general or vehicle insurance. There are also cases where an accident has never occurred, and the vehicles have only been placed onto the road. Still, the majority of such fraud is not planned (*opportunistic fraud*) – an individual only seizes the opportunity arising from the accident and issues exaggerated insurance claims or claims for past damages.

Staged accidents have several common characteristics. They occur in late hours and non-urban areas in order to reduce the probability of witnesses. Drivers are usually younger males, there are many passengers in the vehicles, but never children or elders. The police is always called to the scene to make the subsequent acquisition of means easier. It is also not uncommon that all of the participants have multiple (serious) injuries, when there is almost no damage on the vehicles. Many other suspicious characteristics exist, not mentioned here.

The insurance companies place the most interest in organized groups of fraudsters consisting of drivers, chiropractors, garage mechanics, lawyers, police officers, insurance workers and others. Such groups represent the majority of revenue leakage. Most of the analyses agree that approximately 20% of all insurance claims are in some way fraudulent (various resources). But most of these claims go unnoticed, as fraud investigation is usually done by hand by the domain expert or investigator and is only rarely computer supported. Inappropriate representation of data is also common, making the detection of groups of fraudsters extremely difficult. An expert system approach is thus needed.

Jensen (1997) has observed several technical difficulties in detecting fraud (various domains). Most hold for (automobile) insurance fraud as well. Firstly, only a small portion of accidents or participants is fraudulent (*skewed class distribution*) making them extremely difficult to detect. Next, there is a severe lack of *labeled* data sets as labeling is expensive and time consuming. Besides, due to sensitivity of the domain, there is even a lack of unlabeled data sets. Any approach for detecting such fraud should thus be founded on moderate resources (data sets) in order to be applicable in practice. Fraudsters are very innovative and new types of fraud emerge constantly. Hence, the approach must also be highly adaptable, detecting new types of fraud as soon as they are noticed. Lastly, it holds that fully autonomous detection of automobile insurance fraud is not possible in practice. Final assessment of potential fraud can only be made by the domain expert or investigator, who also determines further actions in resolving it. The approach should also support this investigation process.

Due to everything mentioned above, the set of approaches for detecting such fraud is extremely limited. We propose a novel expert system approach for detection and subsequent investigation

* Corresponding author. Tel.: +386 1 4768 186.
E-mail addresses: lovro.subelj@fri.uni-lj.si (L. Šubelj), stefan.furlan@fri.uni-lj.si (Š. Furlan), marko.bajec@fri.uni-lj.si (M. Bajec).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.07.143

of automobile insurance fraud. The system is focused on detection of groups of collaborating fraudsters, and their connecting accidents (non-opportunistic fraud), and not some isolated fraudulent entities. The latter should be done independently for each particular entity, while in our system, the entities are assessed in a way that considers also the relations between them. This is done with appropriate representation of the domain – networks.

Networks are the most natural representation of any relational domain, allowing formulation of complex relations between entities. They also present the main advantage of our system against other approaches that use a standard *flat data* form. As collaborating fraudsters are usually related to each other in various ways, detection of groups of fraudsters is only possible with appropriate representation of data. Networks also provide clear visualization of the assessment, crucial for the subsequent investigation process.

The system assesses the entities using a novel *Iterative Assessment Algorithm* (*IAA* algorithm), presented in this article. No learning from initial labeled data set is done, the system rather allows simple incorporation of the domain knowledge. This makes it applicable in practice and allows detection of new types of fraud as soon as they are encountered. The system can be used with poor data sets, which is often the case in practice. To simulate realistic conditions, the discussion in the article and evaluation with the prototype system relies only on the data and entities found in the police record of the accident (main entities are participant, vehicle, collision,[1] police officer).

The article makes an in depth description, evaluation and analysis of the proposed system. We pursue the hypothesis that automobile insurance fraud can be detected with such a system and that proper data representation is vital. Main contributions of our work are: (1) a novel expert system approach for the detection of automobile insurance fraud with networks; (2) a benchmarking study, as no expert system approach for detection of groups of automobile insurance fraudsters has yet been reported (to our knowledge); (3) an algorithm for assessment of entities in a relational domain, demanding no labeled data set (*IAA* algorithm); and (4) a framework for detection of groups of fraudsters with networks (applicable in other relational domains).

The rest of the article is organized as follows. In Section 2 we discuss related work and emphasize weaknesses of other proposed approaches. Section 3 presents formal grounds of (social) networks. Next, in Section 4, we introduce the proposed expert system for detecting automobile insurance fraud. The prototype system was evaluated and rigorously analyzed on real world data, description of the data set and obtained results are given in Section 5. Discussion of the results is conducted in Section 6, followed by the conclusion in Section 7.

2. Related work

Our work places in the wide field of fraud detection. Fraud appears in many domains including telecommunications, banking, medicine, e-commerce, general and automobile insurance. Thus a number of expert system approaches for preventing, detecting and investigating fraud have been developed in the past. Researches have proposed using some standard methods of data mining and machine learning, *neural networks*, *fuzzy logic*, *genetic algorithms*, *support vector machines*, (*logistic*) *regression*, *consolidated* (*classification*) *trees*, approaches over *red-flags* or *profiles*, various statistical methods and other methods and approaches (Artis et al., 2002; Bolton and Hand, 2002; Brockett et al., 2002; Estevez et al., 2006; Furlan and Bajec, 2008; Ghosh and Schwartzbard, 1999; Hu et al., 2007; Kirkos et al., 2007; Perez et al., 2005; Quah and Sriganesh, 2008; Rupnik et al., 2007; Sanchez et al., 2009; Viaene et al., 2005; Viaene et al., 2002; Weisberg and Derrig, 1998; Yang and Hwang, 2006). Analyses show that in practice none is significantly better than others (Bolton and Hand, 2002; Viaene et al., 2005). Furthermore, they mainly have three weaknesses. They (1) use inappropriate or inexpressive representation of data; (2) demand a labeled (initial) data set; and (3) are only suitable for larger, richer data sets. It turns out that these are generally a problem when dealing with fraud detection (Jensen, 1997; Phua et al., 2005).

In the narrower sense, our work comes near the approaches from the field of network analysis, that combine intrinsic attributes of entities with their relational attributes. Noble et al. (2003) proposed detecting anomalies in networks with various types of vertices, but they focus on detecting suspicious structures in the network, not vertices (i.e. entities). Besides that, the approach is more appropriate for larger networks. Researchers also proposed detecting anomalies using measures of centrality (Freeman, 1977, 1979), random walks (Sun et al., 2005) and other (Holder and Cook, 2003; Maxion and Tan, 2000), but these approaches mainly rely only on the relational attributes of entities.

Many researchers have investigated the problem of classification in the relational context, following the hypothesis that classification of an entity can be improved by also considering its related entities (inference). Thus many approaches formulating *inference*, *spread* or *propagation* on networks have been developed in various fields of research (Brin and Page, 1998; Domingos and Richardson, 2001; Kleinberg, 1999; Kschischang and Frey, 1998; Lu and Getoor, 2003b; Minka, 2001; Neville and Jensen, 2000). Most of them are based on one of the three most popular (approximate) inference algorithms: *Relaxation Labeling* (RL) (Hummel and Zucker, 1983) from the computer vision community, *Loopy Belief Propagation* (LBP) on loopy (Bayesian) *graphical models* (Kschischang and Frey, 1998) and *Iterative Classification Algorithm* (ICA) from the data mining community (Neville and Jensen, 2000). For the analyses and comparison see (Kempe et al., 2003; Sen and Getoor, 2007).

Researchers have reported good results with these algorithms (Brin and Page, 1998; Kschischang and Frey, 1998; Lu and Getoor, 2003b; Neville and Jensen, 2000), however they mainly address the problem of learning from an (initial) labeled data set (*supervised learning*), or a partially labeled (*semi-supervised learning*) (Lu and Getoor, 2003a), therefore the approaches are generally inappropriate for fraud detection. The algorithm we introduce here, *IAA* algorithm, is almost identical to the *ICA* algorithm, however it was developed with different intentions in mind – to assess the entities when no labeled data set is at hand (and not for improving classification with inference). Furthermore, *IAA* does not address the problem of *classification*, but *ranking*. Thus, in this way, it is actually a simplification of *RL* algorithm, or even Google's *PageRank* (Brin and Page, 1998), still it is not founded on the probability theory like the latter.

We conclude that due to the weaknesses mentioned, most of the proposed approaches are inappropriate for detection of (automobile) insurance fraud. Our approach differs, as it does not demand a labeled data set and is also appropriate for smaller data sets. It represents data with networks, which are one of the most natural representation and allow complex analysis without simplification of data. It should be pointed out that networks, despite their strong foundations and expressive power, have not yet been used for detecting (automobile) insurance fraud (at least according to our knowledge).

3. (Social) networks

Networks are based upon mathematical objects called *graphs*. Informally speaking, graph consists of a collection of points, called

[1] Throughout the article the term collision is used instead of (traffic) accident. The word accident implies there is no one to blame, which contradicts with the article.