Expert Systems with Applications 40 (2013) 4805–4811

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Finding the needle: A risk-based ranking of product listings at online auction sites for non-delivery fraud prediction

V. Almendra

University of Bucharest, Strada Academiei 14, 010014 Bucharest, sector 1, Romania

ARTICLE INFO

Keywords:
Fraud detection
Non-delivery fraud
Boosted trees
E-commerce
Online auction sites
Machine learning
Data collection

ABSTRACT

Non-delivery fraud is a recurring problem at online auction sites: false sellers that list nonexistent products just to receive payments and afterwards disappear, possibly repeating the swindle with another identity. In our work we identified a set of publicly available features related to listings, sellers and product categories, and built a machine learning system for fraud prediction taking into account the high class imbalance of real data and the need to control the false positives rate due to commercial reasons. We tested the proposed system with data collected from a major Brazilian online auction site, obtaining good results on the identification of fraudsters before they strike, even when they had no previous historical information. We also evaluated the contribution of category-related features to fraud detection. Finally, we compared the learning algorithm used (boosted trees) with other state-of-the-art methods.

1. Introduction

Online auction sites like eBay offer unprecedented business possibilities for sellers and buyers through the creation of virtual marketplaces of global reach. Criminals also realized the opportunities opened by such virtual marketplaces. Among the several types of fraudulent behavior that take place in online auction sites, the most frequent one is non-delivery fraud (Gavish & Tucci, 2008; Gregg & Scott, 2008): fake sellers list nonexistent products for sale, receive payments and disappear, possibly reentering the market with a different identity. According to the Internet Crime and Complaint Center (Internet Crime & Complaint Center, 2011), non-delivery fraud is the fourth most reported Internet crime. The challenge faced by site operators is to identify fraudsters *before* they strike, in order to avoid losses due to unpaid taxes, insurance, badmouthing etc. (Chang & Chang, 2011). In other words, for a given product listing they need to *predict* whether or not it will end up being a fraud case. Since online auction sites are huge information systems and all transactions are carried over electronically, a natural approach to the fraud prediction problem is to use machine learning techniques.

In this paper we will present a system for predicting non-delivery fraud that takes as input a set of product listings of an online auction site and outputs for each listing a fraud score, which can be used to analyze listings in decreasing order of risk. It also chooses a risk threshold so as to satisfy the user constraint on the rate of false positives. The proposed system uses a combination of features from product, seller and category, and, unlike other systems in the literature, it depends neither on historical data nor on social networks about the sellers in question, which is an advantage when dealing with fraudsters without reputation. The features we used can be extracted from the public web pages of online auction sites, which means that our system could be implemented by a third party, without the need of internal information. We evaluated the proposed system using data collected from a major Brazilian online auction site.

In Section 2 we will present the context for our research; in Section 3 we will describe the dataset used to validate our approach and will present the selected features; in Section 4 we will explain our proposed system for predicting non-delivery fraud; in Section 5 we will present the experimental results, and in Section 6 we will discuss them.

2. Background

Bolton and Hand (2002) did a comprehensive review regarding statistical fraud detection in several domains: credit card fraud, money laundering, telecommunications fraud, computer intrusion, and scientific fraud. They highlighted some challenges: the high number of cases to be analyzed, the need of fast algorithms, uneven class sizes (class imbalance), uneven misclassification cost, and the problem of false positives. Although they did not mention fraud at online auction sites, these challenges also apply.

There are recently published papers specifically focused on fraud at online auction sites, some from a descriptive perspective (Almendra, 2012; Gavish & Tucci, 2006; Gregg & Scott, 2006), and others aiming fraud prediction (Almendra & Enachescu,

E-mail address: vinicius.almendra@gmail.com

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.02.027

2012; Chang & Chang, 2011; Chau & Faloutsos, 2005; Chiu, Ku, Lie, & Chen, 2011; Maranzato, Pereira, Lago, & Neubert, 2010; Pandit, Chau, Wang, & Faloutsos, 2007; Zhang, Yang, Chu, & Tseng, 2011, 2012). Fraud prediction systems need to tackle the problems of *feature extraction* and *method selection*. Regarding feature extraction, some works relied on public information obtained from online auction sites portals' (Chau & Faloutsos, 2005; Chang & Chang, 2011; Liu, Kaszuba, Nielek, Datta, & Wierzbicki, 2010; Pandit et al., 2007); some used features related to seller past transactions e.g. average product price in the last 15 days (Chau & Faloutsos, 2005; Chang & Chang, 2011; Liu et al., 2010); others used information extracted from the social network surrounding sellers (Pandit et al., 2007); one made use of time-related variations of seller behavior (Chang & Chang, 2011). In our research we included contextual information related to the *category* of the listed products: average price, number of sellers that listed products in the same category, frequency of fraudulent behavior etc. This allowed us to check for example if a listing's price is much below the average. This idea also appeared in another work (Liu et al., 2010), although with much fewer features. There are also works that used internal information of online auction sites (Maranzato et al., 2010; Zhang et al., 2011, Zhang, Yang, & Tseng, 2012), which offers a richer set of features, at the expense of confidentiality restrictions concerning what can be disclosed.

Regarding the methods employed to create classification models, previous works explored several of them: decision trees (Chau & Faloutsos, 2005; Chiu et al., 2011), Markov random fields (Pandit et al., 2007), instance-based learners (Chang & Chang, 2011), logistic regression (Zhang et al., 2011), online probit models (Zhang et al., 2012), Adaptive Neuro-Fuzzy Inference System (Lin, Jheng, & Yu, 2012). The present work uses a variant of boosted trees (Friedman, 2001) as its learning algorithm and compares its performance with several others well-know machine learning techniques.

Another problem related to fraud detection is *class imbalance*: the number of instances of the "positive" class (in our case, fraudulent) is much smaller than the number of instances in the "negative" class (in our case, legitimate). Class imbalance is an obstacle for the use of supervised learning systems in fraud prediction (Bolton & Hand, 2002), since algorithms tend to privilege the prevalent class (in our case, legitimate listings). Some common approaches to solve this problem are undersampling of the majority class, oversampling of the minority class, and SMOTE (Synthetic Minority Over-sampling Technique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Some of the above-mentioned works used undersampling (Chau & Faloutsos, 2005; Chang & Chang, 2011), one uses an unsupervised model (Pandit et al., 2007), others did not state the approach adopted (Maranzato et al., 2010; Zhang et al., 2011, 2012).

3. Dataset description

We already described the dataset used in a previous work (Almendra & Enachescu, 2012). We reproduced the description here for sake of completeness, making some small improvements.

3.1. Data collection

We targeted in our research one specific online auction site, named MercadoLivre (www.mercadolivre.com.br). It is the biggest Brazilian auction site, online since 1999. From now on we will refer to MercadoLivre as MELI. In the whole year of 2011 we crawled daily 11 categories of products where we expected more fraud occurrence, extracting information about 2 million product listings. Using a previously developed methodology (Almendra & Enachescu, 2011), we found 1018 listings with clear signs of non-delivery fraud. Of these 1018, we identified 439 listings about which we had enough information for early fraud prediction. These 439 listings were labeled as *fraudulent listings*. All other listings of active sellers were labeled as *legitimate listings*. Notice that we did not include in our analysis listings of sellers sanctioned by MELI due to other kinds of misbehavior (misrepresentation, fee stacking, unpaid taxes etc.).

3.2. Features for fraud prediction

Our unit of observation was the *product listing*, so our features were also directly or indirectly linked to it. The directly linked features were *price*, *date* (when the listing was published), *product category* and *seller* (MELI's user who owned the listing). We also included information related to the seller: *reputation score*, *account age* (how old the seller account was, in days), and *number of recent transactions*. The values of these features were the ones collected at the day the listing was published in MELI's site (the value of *date* feature), since we wanted to predict fraud *before* transactions took place and before any sign of suspicion. This was possible because we did a longitudinal data collection.

We also included features related to the *product category*, since we expected that fraudsters would not choose randomly which products to list. Product categories specify the type of products, their models, characteristics etc., and sellers have to choose the category in which their products will be listed. Category-related features give us a chance to evaluate a product listing in a wider context. Two roughly equivalent product listings do not necessarily have the same risk of being fraudulent if they belong to very distinct categories.

We used our dataset to obtain aggregated measures about product categories over the entire year of 2011. All listings that shared the same category had the same values for these measures.

Product categories in MELI are organized as a forest, with 23 root nodes and a depth up to 6. Each listing belongs to a specific category *and to all its ancestors*. In other words, each category is a subset of its parent. Fig. 1 exemplifies this structure for one top-level category. The measures related to a category were calculated using the listings specifically belonging to it together with the listings belonging to its descendants.

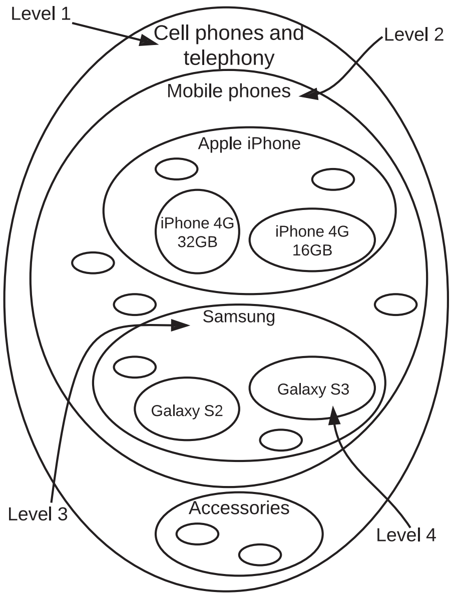

Fig. 1. Excerpt of MELI's category hierarchy as a Venn diagram.