Electronic Commerce Research and Applications 12 (2013) 195–207

Contents lists available at SciVerse ScienceDirect

Electronic Commerce Research and Applications

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

Pricing fraud detection in online shopping malls using a finite mixture model

Kwanho Kim, Yerim Choi, Jonghun Park *

Dept. of Industrial Engineering, Seoul National University, Seoul 151-744, Republic of Korea

ARTICLE INFO

Article history:
Received 7 March 2012
Received in revised form 6 January 2013
Accepted 9 January 2013
Available online 1 February 2013

Keywords:
Pricing fraud
Fraud detection
Online shopping
e-Commerce
Expectation maximization algorithm
Finite mixture model

ABSTRACT

Although pricing fraud is an important issue for improving service quality of online shopping malls, research on automatic fraud detection has been limited. In this paper, we propose an unsupervised learning method based on a finite mixture model to identify pricing frauds. We consider two states, normal and fraud, for each item according to whether an item description is relevant to its price by utilizing the known number of item clusters. Two states of an observed item are modeled as hidden variables, and the proposed models estimate the state by using an expectation maximization (EM) algorithm. Subsequently, we suggest a special case of the proposed model, which is applicable when the number of item clusters is unknown. The experiment results show that the proposed models are more effective in identifying pricing frauds than the existing outlier detection methods. Furthermore, it is presented that utilizing the number of clusters is helpful in facilitating the improvement of pricing fraud detection performances.

1. Introduction

Contemporary online shopping services allow comparison of shopping items in terms of their prices, and have increased price sensitivity of consumers (Alba et al. 1997, Bakos 1997). This phenomenon leads to severe competition among sellers since items that are not attractively priced receive less attention from customers. Accordingly, sellers are compelled to offer items at lower prices than their competitors in order to catch the eyes of customers (Iyer and Pazgal 2003, Kocas 2002).

Such severe competition often causes sellers to set inappropriate prices on their items, resulting in pricing fraud that attempts to deceive customers by offering seemingly lower prices on items as compared to normal market prices (Gregg and Scott 2006). Recent social commerce and online auction services such as eBay exacerbate the problem, since sellers are able to sell their items without being investigated for possible pricing fraud in those types of services (Dekleva 2000, Chua and Wareham 2004, Gavish and Tucci 2006).

In practice, however, service providers still heavily rely on manual investigation for identifying pricing frauds, which is costly and time consuming (Zhang et al. 2012). Pricing fraud has been usually detected through examining some fraudulent seller behaviors reported in online auctions such as misrepresentation (Ba et al. 2003, Zacharia et al. 2000), fee stacking (Chua and Wareham 2004), and price shilling (Kauffman and Wood 2005, Dong et al. 2012), which result in the irrelevance between an item's price and its description. Gregg and Scott (2006), Gregg and Scott (2008) reported that these behaviors account for approximately 27.5–32.3% of fraudulent activities in online auction services.

Regardless of whether customers actually purchase fraudulently priced items, such items become a subject of major concern for both customers and service providers because pricing fraud significantly impairs service quality. Customers attempting to purchase a fraud item may be eventually required to pay additional charges or become embarrassed due to the difference between an item's final price including additional charges and the initial price presented by a seller. On the other hand, eliminating pricing fraud is crucial for a service provider to improve customer trust, since customers repeatedly encountering pricing fraud become distrustful of item prices, leading to a decreased number of returning customers and lower turnover eventually (Gavish and Tucci 2008).

As an example, we consider the item description of a camera package, "Cannon EOS 600D 1855 mm Genuine Full Package", priced at 790,000 KRW (Korean won) from an online shopping service, Auction (http://www.auction.co.kr). This item would appeal to customers since its price is quite low compared to those of other items with the same purchase options. However, during checkout, the customer is requested to pay an additional 230,000 KRW for the camera lens, originally presented as an included component in the item description. With this extra charge, the item loses its price advantage.

Identification of pricing fraud is a challenging problem owing to the following two major characteristics: First, actual data on fraudulently priced items are often unavailable due to the significant time and cost required for manual fraud detection, and this is

* Corresponding author.
E-mail address: jonghun@snu.ac.kr (J. Park).

1567-4223/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.elerap.2013.01.001

particularly true for small and medium-sized online shopping malls. Second, item price changes dynamically due to the frequent price updates of shopping items by sellers (Nakamura 1999, Tang and Xing 2001). In the dataset considered in this paper, about 10% of item prices were updated daily, and an item price changed at least every five days, implying that normal price range of an item varies over time.

Accordingly, the existing fraud detection approaches based on the notion of similarities against the previously observed normal and fraud items cannot be directly applied to the considered pricing fraud problem. Specifically, while supervised approaches such as neural networks (Aleskerov et al. 1997, Dorronsoro et al. 1997) and rule-based systems (Brause et al. 1999) yielded satisfactory results in various applications including the fraud detection, they are not applicable to the detection of pricing fraud because of their reliance on historical data to estimate model parameters. Moreover, unsupervised approaches such as clustering based methods (Bolton and Hand 2001) and outlier detection methods (Eskin 2010, Liu et al. 2003, Sabuncu et al. 2010, Yang et al. 2009) may not be effective for the considered problem, since they cannot accommodate the fact that the relevance between the item features such as item description and price varies depending on whether a seller's behavior is fraudulent or not.

Motivated by the challenges mentioned above, we propose novel models designed to automatically detect pricing fraud of items in online shopping malls by considering the relevance among the item features depending on a seller's behavior. The idea of fraud identification based on the dependency among features is not new and was originally proposed by Huang et al. (2003) and Hu and Panda (2005) in which cross-feature analysis was proposed to examine whether or not the features of instances are relevant to each other by using the frauds detected previously. Our approach utilizes this cross-feature analysis, but does not require actual fraud data for constructing a model.

The proposed models base on a finite mixture model widely known for its effectiveness and flexibility (McLachlan and Peel 2000). In contrast to the supervised learning methods, such as support vector machines and decision trees, the finite mixture model is applicable even when labeled data are not available, and allows explicit modeling of the dependency between item description and price (Everitt and Hand 1981).

In the proposed approach, two states, namely normal and fraud states, are defined for each item in a set of observed items according to a seller's pricing behavior. These two item states are modeled by using three hidden variables that indicate the corresponding cluster of an item description, the cluster of item price, and the dependency between them. The proposed models then estimate the item state based on the dependency between the item description and price, and they determine the clusters of the item description and the price for each item through investigating the possible combinations of those clusters.

The proposed approach further attempts to more precisely infer fraudulent items by utilizing the known number of item clusters. An item cluster represents a subset of items in an itemset, a set of items from the same item type, specified based on available purchase options. The number of item clusters for each itemset can be obtained easily from many online shopping malls, unlike the other information such as sellers' reputation data and labels indicating whether or not an item is fraudulently priced, which are costly to obtain. For instance, USB drive products are usually grouped into multiple item clusters in terms of their capacity: "4 GB", "8 GB", and "16 GB" in online shopping malls.

The paper is organized as follows. Section 2 reviews related studies to this research. In Section 3, we define the proposed model, called pricing fraud detection model with the known number of item clusters (PDMC), and we also examine its special case, called pricing fraud detection model (PDM). In Section 4, we present the experiment results to show the effectiveness of the proposed models by using a real-world dataset. In Section 5, we discuss the implications and limitations of the proposed models. Finally, the conclusions are presented in Section 6.

2. Literature review

During the last two decades, there have been many research results on fraud detection which can be broadly grouped into three categories according to the approaches taken. First, distance-based methods that focus on the distance between an instance and the other normal or fraud instances observed previously, have been widely applied to identify fraudulent credit card usage (Aleskerov et al. 1997, Dorronsoro et al. 1997, Brause et al. 1999), insider trading (Arning et al. 1996), and fraudulent mobile phone use (Taniguchi et al. 1998, Phua et al. 2004). Unfortunately, they require a sufficient number of labeled normal and fraud instances for identifying frauds satisfactorily.

Second, anomaly based methods aim to find anomalies from instances without requiring previously labeled data. They measure the degree of fraud for an instance based on the distance between the instance and a group of frequently observed instances. Based on this approach, some results for fraud detection in mobile phone usage and insurance claim have been reported (He et al. 2003). While this approach does not require labeled data, it involves manual intervention for deciding model parameters and cannot consider the possible dependency between item features.

Finally, sequence-based methods have been developed to identify unusual changes in behavior over time compared to those observed in the past. These methods have been widely used for detecting fraudulent credit card usage (Bolton and Hand 2001), insider trading (Donoho 2004), insurance claim fraud (Fawcett and Provost 1999), and fraudulent mobile phone use (Aggarwal 2005).

Along with the research results that have successfully addressed the various types of frauds, there is also a line of research specifically focused on detecting frauds related to shopping items. Some studies have emphasized the effective use of seller reputation information obtained from rating systems and social networks. Ku et al. (2007) proposed a decision tree based method that finds potential fraudulent sellers in online auction services by analyzing seller reputations obtained from customer comments in social networks. Zhang et al. (2008) also presented a fraud classification method using a Markov random field of network level data obtained from sellers' historical reputation records.

In addition, a content analysis method that utilizes complaint data and ratings of sellers in a reputation system (Gregg and Scott 2006, 2008) has been reported. Chang and Chang (2010) and You et al. (2011) also developed similar approaches using the data obtained from reputation systems. Yet, these approaches require seller reputation data which are costly to obtain, and they are reactive rather than being proactive from the viewpoint of pricing fraud detection since customer comments and seller ratings become meaningful only after frauds have actually occurred.

A more comprehensive approach that employs both seller information and transaction logs to detect frauds has been reported in Chau and Faloutsos (2007) where a decision-tree-based method that requires seller profiles and transaction logs as well as labeled data was proposed to detect fraudulent offers by analyzing selling patterns. More recently, Chang and Chang (2012) presented an approach for early fraud detection by using various combinations of item features such as prices, price changes, and transaction logs.

Conventional fraud detection methods focus on identifying frauds mainly based on the distance between items or the degree of anomaly without considering the relevance between the