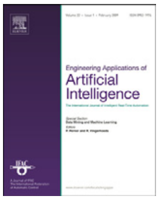
Engineering Applications of Artificial Intelligence 24 (2011) 182–194

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



A data mining framework for detecting subscription fraud in telecommunication

Hamid Farvaresh, Mohammad Mehdi Sepehri*

Department of Industrial Engineering, Tarbiat Modares University, 1411713114 Tehran, Iran

ARTICLE INFO

Article history:
Received 9 September 2008
Received in revised form
9 April 2010
Accepted 26 May 2010
Available online 19 June 2010

Keywords:
Fraud detection
Data mining
Neural networks
Decision tree
Support vector machines
Ensembles
Telecommunication.

ABSTRACT

Service providing companies including telecommunication companies often receive substantial damage from customers' fraudulent behaviors. One of the common types of fraud is subscription fraud in which usage type is in contradiction with subscription type. This study aimed at identifying customers' subscription fraud by employing data mining techniques and adopting knowledge discovery process. To this end, a hybrid approach consisting of preprocessing, clustering, and classification phases was applied, and appropriate tools were employed commensurate to each phase. Specifically, in the clustering phase SOM and K-means were combined, and in the classification phase decision tree (C4.5), neural networks, and support vector machines as single classifiers and bagging, boosting, stacking, majority and consensus voting as ensembles were examined. In addition to using clustering to identify outlier cases, it was also possible – by defining new features – to maintain the results of clustering phase for the classification phase. This, in turn, contributed to better classification results. A real dataset provided by Telecommunication Company of Tehran was applied to demonstrate the effectiveness of the proposed method. The efficient use of synergy among these techniques significantly increased prediction accuracy. The performance of all single and ensemble classifiers is evaluated based on various metrics and compared by statistical tests. The results showed that support vector machines among single classifiers and boosted trees among all classifiers have the best performance in terms of various metrics. The research findings show that the proposed model has a high accuracy, and the resulting outcomes are significant both theoretically and practically.

1. Introduction

Telecommunication businesses are producing and storing a huge amount of data all over the world. These data are very interesting for data mining applications. The main feature of these great databases is their extraordinary size. More than 300 million records per day, for example, are stored in AT&T solely for long-distance calls (Cortes and Pregibon, 2001). Although these companies own a great source of information, only few of them are aware of the hidden knowledge of these databases. Thus, they do not use it frequently in their decision making processes.

A challenge that not only telecommunication companies but also other service institutions such as banks, water and energy suppliers, and credit companies confront is customers' fraud detection. Fraud in telecom services causes a substantial loss of annual revenue for many telecommunication companies throughout the world (Paredes, 2005; Xing and Girolami, 2007). There are different types of fraud in the telecommunication business (Shawe-Taylor et al., 1999). Shawe-Taylor et al. (2000) present six different fraud types: subscription fraud, the manipulation of Private Branch Exchange (PBX) facilities or dial through fraud, free phone fraud, premium rate service fraud, handset theft, and roaming fraud.

A common type of fraud is subscription fraud (Estevez et al., 2006). Many companies offer lower tariffs for residential subscribers than for commercial ones. So customers may ask for residential subscription, but use it for commercial purposes. In wireline telephone service, identifying the subscription fraudulent customers is possible by checking the installation and usage place. However, identifying all fraudulent customers through checking all residential customers in companies like Telecommunication Company of Tehran (TCI), which has millions of residential customers, needs a lot of money and time. Therefore, reducing the number of customers to be checked is very demanding. This study intends to propose a method to detect different patterns of residential and commercial subscribers' behaviors based on their call detail recording (CDR) and bills' data in order to differentiate residential subscription, which have a behavior similar to fraudulent customers. We have tried to recognize the true subscription type with the highest accuracy. Detecting subscription fraud can prevent a great part of telecommunication income loss.

* Corresponding author. Tel.: +98 21 8288 3379.
E-mail addresses: mehdi.sepehri@modares.ac.ir, mehdi.sepehri@gmail.com (M.M. Sepehri).

0952-1976/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2010.05.009

The remaining of this paper is organized as follows: Section 2 reviews the previous literature on the techniques for customers' fraud detection. The proposed method is then described in Section 3. In Section 4, a real dataset provided by Telecommunication Company of Iran (hereinafter called TCI), is applied as a case study to demonstrate the effectiveness of the proposed method. Finally, concluding remarks are offered in Section 5.

2. Literature review

It is a very common interest among telecommunication companies to extract an accurate profile for each subscriber based on his/her CDR patterns. Subscribers' profiles not only are useful for detecting abnormal behaviors but also mainly used for marketing purposes and customer relationship management (CRM) (Sohn and Kim, 2008). These profiles are based on either CDR (e.g., number of calls, call duration, call type) or subscriber demographic properties (e.g., age, gender, region) or both. Generally speaking, customer fraud detection techniques divide customers into two groups of normal and fraudulent customers. Many researchers studied customer fraud detection in telecommunication companies using data mining techniques (Barson et al., 1996; Fawcett and Provost, 1997; Shawe-Taylor et al., 1999). Detecting insolvent customers (Ezawa, 1996; Daskalaki et al., 2003), detecting subscription fraud (Cahill et al., 2002; Cortes et al., 2003; Estevez et al., 2006), and detecting fax lines from telephone lines (Kaplan et al., 1999) are some examples of recent studies.

Several techniques have been proposed to create customer's profile. Fraud detection in telecommunication business is mostly done for mobile services (Buschkes et al., 1998; Gosset and Hyland, 1999; Burge and Shawe-Taylor, 2001). Methods like rule mining (Fawcett and Provost, 1997; Adomavicious and Tuzhilin, 1999), clustering (Oh and Lee, 2003), Bayesian network (Buschkes et al., 1998), neural network (Manikopoulos and Papavassilliou, 2002; Daskalaki et al., 2003), latent Dirichlet allocation (Xing and Girolami, 2007), and decision tree (Daskalaki et al., 2003) are some examples in this context. A comprehensive survey of data mining techniques applied to various fraud detection problems was presented in Phua et al. (2005).

Roughly speaking, detecting fraudulent subscribers is a classification problem that may be solved by various data mining techniques. These techniques vary in terms of statistical techniques (e.g., regression family techniques), artificial intelligence techniques (e.g., decision trees, neural networks), dimension reduction method (e.g., PCA, MDS), number of features included in the model, as well as feature-selection method (e.g., theory versus stepwise selection). In any case, a classifier should classify each customer into one of the two classes of normal or fraudulent customers. However, the major challenges in detecting telecommunication fraud are: (1) to deal with the high volume of call traffic in an efficient manner; (2) to be effective at identifying small percentage calls which are fraudulent; and (3) to be done at a reasonably low cost (Kou et al., 2004; Xing and Girolami, 2007).

In the next section an overview of the applied tools and techniques is presented. It should be noted that due to space limitation technical and mathematical details of the methods are not presented here, and interested readers are referred to appropriate references.

2.1. Discriminant analysis (DA) and logistic regression (LR)

Statistical methods (parametric and non-parametric) along with artificial intelligence and machine learning have been widely used to make classifiers (Desai et al., 1996; Thomas, 2000; West, 2000). Linear discriminant analysis (LDA) is the oldest and most common statistical tool in handling classification problems (Lee et al., 1999) that has been employed in fraud detection and credit scoring (Desai et al., 1996; Daskalaki et al., 2003; Lee et al., 2006). Reichert et al. (1983) mentioned two disadvantages for LDA: assuming (1) multivariate normality of metric independent variables and (2) homogeneous covariance matrix between categorical dependent variable (classes). However, since the independent variables are a mixture of categorical and continuous variables, the multivariate normality assumption will not hold. LR is another statistical method which is used in credit scoring models (Laitinen and Laitinen, 2000; Westgaard and van der Wijst, 2001). The advantage of LR in comparison with LDA is that LR does not make any assumption about the distribution of the independent variables. Furthermore, the result of classification is not in the YES/NO format, but is an estimated probability of each observation belonging to a given class, and is easy to interpret (Timm, 2002). LR, however, has also some restricting assumptions which may not be always true. Homogeneous variance matrix between classes is an example in this regard (McCulloch and Searle, 2001). The interested reader is referred to several relevant references (Sharma, 1996; Timm, 2002; Izenman, 2008).

2.2. Neural networks (NN)

Recently, NN has been widely used due to its ability for modeling complex and non-linear models, and also not having any strict limitations and rigorous assumption for the type of input data (Stern, 1996; Anderson and Rosenfeld, 1998). As a negative aspect, on the other hand, one can mention long learning time, overfitting error, and black box characteristics of NN (Bishop, 1995; Hippert et al., 2005). NN has also been used for telecommunication and bank customers' credit scoring problem (West, 2000; Lee et al., 2002; Mahlhotra and Malhotra, 2003; Hsieh, 2005). For a comprehensive account of neural networks refer to (Bishop, 1995).

2.3. Decision tree (DT)

C4.5 decision tree learning is one of the most widely used and practical methods for inductive inference. It is a method for classification that is robust to noisy data and capable of learning disjunctive expressions. Decision tree learning is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree. Learned trees can also be represented as sets of if–then rules to improve human readability. Among rule based methods, DT models including CART, ID3, and C4.5 have been used for customer classification, fraud detection, and customers' credit scoring problems more than other methods (Rosset et al., 1999; Shao et al., 2002; Daskalaki et al., 2003; Lee et al., 2006). ID3 has been less used because of its categorical property. C4.5 has been recently introduced as a successful tool for customer classification due to its ability to handle continuous variables (Wei and Chiu, 2002; Daskalaki et al., 2003; Hung et al., 2006; Chung and Suh, 2009). An advantage of DT is that it does not have any limitations about data type, and, moreover, it represents the results in an understandable format (Ong et al., 2005). Nevertheless, DT and other rule based methods do not guarantee to cover the whole variable space. So it is possible to have a new instance which cannot be classified by any of the available rules. For more on decision tree (C4.5) see Quinlan (1993).