Decision Support Systems 50 (2011) 602–613

Contents lists available at ScienceDirect

Decision Support Systems

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

Data mining for credit card fraud: A comparative study

Siddhartha Bhattacharyya [a,*], Sanjeev Jha [b,1], Kurian Tharakunnel [c], J. Christopher Westland [d,2]

[a] Department of Information and Decision Sciences (MC 294), College of Business Administration, University of Illinois, Chicago, 601 South Morgan Street, Chicago, Illinois 60607-7124, USA
[b] Department of Decision Sciences, Whittemore School of Business and Economics, University of New Hampshire, McConnell Hall, Durham, New Hampshire 03824-3593, USA
[c] Tabor School of Business, Millikin University, 1184 West Main Street, Decatur, IL 62522, USA
[d] Department of Information & Decision Sciences (MC 294), College of Business Administration, University of Illinois, Chicago, 601 S. Morgan Street, Chicago, IL 60607-7124, USA

ARTICLE INFO

Available online 18 August 2010

Keywords:
Credit card fraud detection
Data mining
Logistic regression

ABSTRACT

Credit card fraud is a serious and growing problem. While predictive models for credit card fraud detection are in active use in practice, reported studies on the use of data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. This paper evaluates two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study is based on real-life data of transactions from an international credit card operation.

1. Introduction

Billions of dollars are lost annually due to credit card fraud [12,14]. The 10th annual online fraud report by CyberSource shows that although the percentage loss of revenues has been a steady 1.4% of online payments for the last three years (2006 to 2008), the actual amount has gone up due to growth in online sales [17]. The estimated loss due to online fraud is $4 billion for 2008, an increase of 11% on the 2007 loss of $3.6 billion [32]. With the growth in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud, and 70% of U.S. consumers are noted to be significantly concerned about identity fraud [35]. Additionally, credit card fraud has broader ramifications, as such fraud helps fund organized crime, international narcotics trafficking, and even terrorist financing [20,35]. Over the years, along with the evolution of fraud detection methods, perpetrators of fraud have also been evolving their fraud practices to avoid detection [3]. Therefore, credit card fraud detection methods need constant innovation. In this study, we evaluate two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study is based on real-life data of transactions from an international credit card operation.

Statistical fraud detection methods have been divided into two broad categories: *supervised and unsupervised* [3]. In supervised fraud detection methods, models are estimated based on the samples of fraudulent and legitimate transactions, to classify new transactions as fraudulent or legitimate. In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both these fraud detection methods predict the probability of fraud in any given transaction.

Predictive models for credit card fraud detection are in active use in practice [21]. Considering the profusion of data mining techniques and applications in recent years, however, there have been relatively few reported studies of data mining for credit card fraud detection. Among these, most papers have examined neural networks [1,5,19,22], not surprising, given their popularity in the 1990s. A summary of these is given in [28], which reviews analytic techniques for general fraud detection, including credit card fraud. Other techniques reported for credit card fraud detection include case based reasoning [48] and more recently, hidden Markov models [45]. A recent paper [49] evaluates several techniques, including support vector machines and random forests for predicting credit card fraud. Their study focuses on the impact of aggregating transaction level data on fraud prediction performance. It examines aggregation over different time periods on two real-life datasets and finds that aggregation can be advantageous, with aggregation period length being an important factor. Aggregation was found to be especially effective with random forests. Random forests were noted to show better performance in relation to the other techniques, though logistic regression and support vector machines also performed well.

Support vector machines and random forests are sophisticated data mining techniques which have been noted in recent years to show superior performance across different applications [30,38,46,49]. The choice of these two techniques, together with logistic regression, for this study is based on their accessibility for practitioners, ease of use, and noted performance advantages in the literature. SVMs are statistical learning techniques, with strong

* Corresponding author. Tel.: +1 312 996 8794; fax: +1 312 413 0385.
E-mail addresses: sidb@uic.edu (S. Bhattacharyya), sanjeev.jha@unh.edu (S. Jha), ktharakunnel@millikin.edu (K. Tharakunnel), westland@uic.edu (J.C. Westland).
[1] Tel.: +1 603 862 0314; fax: +1 603 862 3383.
[2] Tel.: +1 312 996 2323; fax: +1 312 413 0385.

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.08.008

theoretical foundation and successful application in a range of problems [16]. They are closely related to neural networks, and through use of kernel functions, can be considered an alternate way to obtain neural network classifiers. Rather than minimizing empirical error on training data, SVMs seek to minimize an upper bound on the generalization error. As compared with techniques like neural networks which are prone to local minima, overfitting and noise, SVMs can obtain global solutions with good generalization error. They are more convenient in application, with model selection built into the optimization procedure, and have also been found to outperform neural networks in classification problems [34]. Appropriate parameter selection is, however, important to obtain good results with SVM.

Single decision tree models, though popular in data mining application for their simplicity and ease of use, can have instability and reliability issues. Ensemble methods provide a way to address such problems with individual classifiers and obtain good generalization performance. Various ensemble techniques have been developed, including mixture of experts, classifier combination, bagging, boosting, stacked generalization and stochastic gradient boosting (see [29] and [40] for a review for these). For decision trees, the random subspace method considers a subset of attributes at each node to obtain a set of trees. Random forests [6] combine the random subspace method with bagging to build an ensemble of decision trees. They are simple to use, with two easily set parameters, and with excellent reported performance noted as the ensemble method of choice for decision trees [18]. They are also computationally efficient and robust to noise. Various studies have found random forests to perform favorably in comparison with support vector machine and other current techniques [10,34].

The third technique included in this study is logistic regression. It is well-understood, easy to use, and remains one of the most commonly used for data-mining in practice. It thus provides a useful baseline for comparing performance of newer methods.

Supervised learning methods for fraud detection face two challenges. The first is of unbalanced class sizes of legitimate and fraudulent transactions, with legitimate transactions far outnumbering fraudulent ones. For model development, some form of sampling among the two classes is typically used to obtain training data with reasonable class distributions. Various sampling approaches have been proposed in the literature, with random oversampling of minority class cases and random undersampling of majority class cases being the simplest and most common in use; others include directed sampling, sampling with generation of artificial examples of the minority class, and cluster-based sampling [13]. A recent experimental study of various sampling procedures used with different learning algorithms [25] found performance of sampling techniques to vary with learning algorithm used, and also with respect to performance measures. The paper also found that simpler techniques like random over and undersampling generally perform better, and noted very good overall performance of random undersampling. Random undersampling is preferred to oversampling, especially with large data. The extent of sampling for best performance needs to be experimentally determined. In this study, we vary the proportion of fraud to non-fraud cases in the training data using random undersampling, and examine its impact in relation to the three learning techniques and considering different performance measures.

The second problem in developing supervised models for fraud can arise from potentially undetected fraud transactions, leading to mislabeled cases in the data to be used for building the model. For the purpose of this study, fraudulent transactions are those specifically identified by the institutional auditors as those that caused an unlawful transfer of funds from the bank sponsoring the credit cards. These transactions were observed to be fraudulent ex post. Our study is based on real-life data of transactions from an international credit card operation. The transaction data is aggregated to create various derived attributes.

The remainder of the paper is organized as follows. Section 2 proves some background on credit card fraud. The next section describes the three data mining techniques employed in this study. In Section 4 we discuss the dataset source, primary attributes, and creation of derived attributes using primary attributes. Subsequently, we discuss the experimental set up and performance measures used in our comparative study. Section 6 presents our results and the final section contains a discussion on findings and issues for further research.

2. Credit card fraud

Credit card fraud is essentially of two types: application and behavioral fraud [3]. Application fraud is where fraudsters obtaining new cards from issuing companies using false information or other people's information. Behavioral fraud can be of four types: mail theft, stolen/lost card, counterfeit card and 'card holder not present' fraud. Mail theft fraud occurs when fraudsters intercept credit cards in mail before they reach cardholders or pilfer personal information from bank and credit card statements [8]. Stolen/lost card fraud happens when fraudsters get hold of credit cards through theft of purse/wallet or gain access to lost cards. However, with the increase in usage of online transactions, there has been a significant rise in counterfeit card and 'card holder not present' fraud. In both of these two types of fraud, credit card details are obtained without the knowledge of card holders and then either counterfeit cards are made or the information is used to conduct 'card holder not present' transactions, i.e. through mail, phone, or the Internet. Card holders information is obtained through a variety of ways, such as employees stealing information through unauthorized 'swipers', 'phishing' scams, or through intrusion into company computer networks. In the case of 'card holder not present' fraud, credit cards details are used remotely to conduct fraudulent transactions.

The evolution of credit card fraud over the years is chronicled in [50]. In the 1970s, stolen cards and forgery were the most prevalent type of credit card fraud, where physical cards were stolen and used. Later, mail-order/phone-order became common in the '80s and '90s. Online fraud has transferred more recently to the Internet, which provides the anonymity, reach, and speed to commit fraud across the world. It is no longer the case of a lone perpetrator taking advantage of technology, but of well-developed organized perpetrator communities constantly evolving their techniques.

Boltan and Hand [4] note a dearth of published literature on credit card fraud detection, which makes exchange of ideas difficult and holds back potential innovation in fraud detection. On one hand academicians have difficulty in getting credit card transactions datasets, thereby impeding research, while on the other hand, not much of the detection techniques get discussed in public lest fraudsters gain knowledge and evade detection. A good discussion on the issues and challenges in fraud detection research is provided in [4] and [42].

Credit card transaction databases usually have a mix of numerical and categorical attributes. Transaction amount is the typical numerical attribute, and categorical attributes are those like merchant code, merchant name, date of transaction etc. Some of these categorical variables can, depending on the dataset, have hundreds and thousands of categories. This mix of few numerical and large categorical attributes have spawned the use of a variety of statistical, machine learning, and data mining tools [4]. We faced the challenge of making intelligent use of numerical and categorical attributes in this study. Several new attributes were created by aggregating information in card holders' transactions over specific time periods. We discuss the creation of such derived attributes in more detail in Section 4 of this paper.

Another issue, as noted by Provost [42], is that the value of fraud detection is a function of time. The quicker a fraud gets detected, the greater the avoidable loss. However, most fraud detection techniques need history of card holders' behavior for estimating models. Past