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