Data Mining for Fraud Detection System

Roberto Marmo
University of Pavia, Italy

INTRODUCTION

As a conseguence of expansion of modern technology, the number and scenario of fraud are increasing dramatically. Therefore, the reputation blemish and losses caused are primary motivations for technologies and methodologies for fraud detection that have been applied successfully in some economic activities. The detection involves monitoring the behavior of users based on huge data sets such as the logged data and user behavior.

The aim of this contribution is to show some data mining techniques for fraud detection and prevention with applications in credit card and telecommunications, within a business of mining the data to achieve higher cost savings, and also in the interests of determining potential legal evidence.

The problem is very difficult because fraudsters takes many different forms and are adaptive, so they will usually look for ways to avoid every security measures.

BACKGROUND

The economic operations under security control can be classified into the class of genuine and into the class of fraudulent. A fraud is a criminal deception, use of false representations to obtain an unjust advantage, or to injure the rights and interests of another. The fraud is prevalent in insurance, credit card, telecommunications, health care, finance, etc. Diversity of fraud regards organisations, governments, and individuals such as external parties, internal employees, customers, service providers and suppliers.

It is important to analyze in detail the fraud scenario in order to establish: what is the fraudulent and normal behavior and what separates one individual from another, the degree of available knowledge about known fraud, the kind of available data exemplifying, types of fraud offenders and their modus operandi over time. It is difficult to provide precise estimates since some fraud may never be detected, and the operators are reluctant to reveal fraud losses due to show an appearance of reliability and security in business operations and to avoid reputation blemish.

It is necessary to take into account the cost of the fraud detection and the cost of fraudulent behavior, because stopping a fraud of few dollars can require a very expensive system. This is possible by introducing a decision layer on top of the system in order to decide the action taking into account factors like the amount of transaction and the risk associated to user doing the transaction.

The development of new detections methods is more difficult due to the severe limitation on privacy and on exchange of ideas. Moreover, data sets are not available and results are often not disclosed to the public.

The planning audit strategies is a posteriori fraud detection problem with prevention purpose of analyzing historical audit data and constructing models of planning effectively future audits. An application is fiscal and insurance domain, where audits are intended to detect tax evasion and fraudulent claims. A case study is presented by Bonchi (1999) which illustrates how techniques based on classification can be used to support the task of planning audit strategies.

The fraud detection methods in online auction (Shah, 2002) are based on statistical methods and association analysis in order to detect shilling, that occurs when the seller tries to hike up the prices in auction by placing buy bids under distinct aliases or through associates.

Apart fraud, the detection efforts may be further motivated by the need to understand the behavior of customers to enable provision of matching services and to improve operations.

DATA MINING APPROACH

Data mining analyzes the huge volumes of transactions and billing data and seeks out patterns, trends and clusters that reveal fraud. The main steps for implementing this approach for fraud detection within a business organization are:

1. Analyze the fraud objectives and the potential fraudsters, in order to converting them into data mining objectives;
2. Data collection and understanding;
3. Data cleaning and preparation for the algorithms;
4. Experiment design;
5. Evaluation results in order to review the process.

Relevant technical problems are due to:

1. Imperfect data not collected for purpose of data mining, so they are inaccurate, incomplete, and irrelevant data attributes;
2. Highly skewed data, there are many more legitimate than fraudulent examples, so by predicting all examples to be legal a very high success rate is achieved without detecting any fraud;
3. Higher chances of overfitting, that occurs when model high accuracy arises from fitting patterns in the training set that are not statistically reliable and not available in the score set.

To handle with skewed data the training set is divided into pieces where the distribution is less skewed (Chan, 1998).

A typical detection approach consists in outlier detection where the non-fraudulent behavior is assumed as normal and identify outliers that fall far outside the expected range should be evaluated more closely. Statistic techniques used for this approach are:

1. Predict and Classify
   - Regression algorithms: neural networks, CART, Regression, GLM;
   - Classification algorithms (predict symbolic outcome): CART, logistic regression;
2. Group and Find Associations
   - Clustering/Grouping algorithms: K-means, Kohonen, Factor analysis;
   - Association algorithms: GRI, Capri Sequence.

Many existing fraud detection systems operate by: supervised approaches on labelled data, hybrid approaches on labelled data, semi-supervised approaches with legal (non-fraud) data, unsupervised approaches with unlabelled data (Phua, 2005).

For a pattern recognition problem requiring great flexibility, adaptivity, and speed, neural networks techniques are suitable and the computational immunological systems, inspired by human immune system, might prove even more effective than neural networks in rooting out e-commerce fraud. (Weatherford, 2002). Unsupervised neural networks can mainly be used because they act on unlabelled data in order to extract an efficient internal representation of the data distribution structure.

The relational approach (Kovalerchuk, 2000) is applicables for discovering financial fraud, because overcome difficulties of traditional data mining in discovering patterns having only few relevant events in irrelevant data and insufficient statistics of relevant data.

The choice of three classification algorithms and one hybrid meta-learning was introduced by Phua (2004), to process the sampled data partitions, combined with straightforward cost model to evaluate the classifiers, so the best mix of classifiers can be picked.

Visualization techniques are based on the human capacity in pattern recognition in order to detect anomalies and are provided with real-time data. A machine-based detection method is static, the human visual system is dynamic and can easily adapt to the typical ever-changing techniques of the fraudsters. Visual data mining is a data mining approach that combine human detection and statistical analysis for greater computational capacity, is developed by building a user interface to manipulate the visual representation of data in fraud analysis.

Service providers use performance metric like the detection rate, false alarm rate, average time to detection after fraud starts, and average number of fraud or minutes until detection. As first step it is important to define a specific metric considering that misclassification costs can differ in each data set and can change over time. The false alarm rate is the percentage of legitimate that are incorrectly identified as fraudulent; fraud catching rate (or true positive rate or detection accuracy rate) is the percentage of transactions that are correctly identified as fraudulent; false negative rate is the percentage of transactions that are incorrectly identified as legitimate. The objective of this detection is to maximize correct fraud predictions and maintain incorrect predictions at an acceptable level. A realistic objective consists on balance of the performance criteria. A false negative error is usually more costly than