Expert Systems with Applications 38 (2011) 13057–13063

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Detecting credit card fraud by genetic algorithm and scatter search

Ekrem Duman [a,*], M. Hamdi Ozcelik [b]

[a] Dogus University, Industrial Engineering Department, Istanbul, Turkey
[b] Yapi Kredi Bankasi, IT Department, Istanbul, Turkey

ARTICLE INFO

Keywords:
Fraud
Credit cards
Genetic algorithms
Scatter search
Optimization

ABSTRACT

In this study we develop a method which improves a credit card fraud detection solution currently being used in a bank. With this solution each transaction is scored and based on these scores the transactions are classified as fraudulent or legitimate. In fraud detection solutions the typical objective is to minimize the wrongly classified number of transactions. However, in reality, wrong classification of each transaction do not have the same effect in that if a card is in the hand of fraudsters its whole available limit is used up. Thus, the misclassification cost should be taken as the available limit of the card. This is what we aim at minimizing in this study. As for the solution method, we suggest a novel combination of the two well known meta-heuristic approaches, namely the genetic algorithms and the scatter search. The method is applied to real data and very successful results are obtained compared to current practice.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

This study is motivated from an industrial consultancy project. Our industrial partner (a major bank in Turkey) has been using an internally developed credit card fraud detection solution for some years. Although that solution has been regarded as successful, the bank authorities thought that it can further be improved due to two expectations/reasons. First, the weights of the parameters used could be better adjusted using the recent card usage behaviors and frauds happened. Second, it has been understood that a good solution is not necessarily the one detecting many frauds but the one detecting frauds maybe fewer in number but larger in risk.

Fraud can be defined as the illegal usage of any system or good. Correspondingly the legal activities can be named as legitimate. We can face with fraud in a variety of different domains including banking, insurance, telecommunications, health care and public services. In banking, frauds can be observed in the use of credit cards, debit cards, internet banking accounts and call center (telephone banking). Money laundering and personnel fraud are the other banking related fraud types. The losses due to fraud sum up to huge amounts and it is a major threat to the legal economy. Inherited to its importance it has attracted the interest of many scientists. During the last 10 years (1999–2009) 1361 articles are found to be published according to the ISI Web of Knowledge data when a search with the keyword "fraud" is made.

In this study we are concerned only with the credit card frauds. When we analyzed the data of our industrial partner and several other banks we observe that only several out of 100,000 transactions are fraudulent transactions. The rest are legitimate. This extremely high imbalance between the two classes makes the fraud detection a challenging task.

Fraud detection has been usually seen as a data mining problem where the objective is to correctly classify the transactions as legitimate or fraudulent. For classification problems many performance measures are defined most of which are related to the correct number of cases classified correctly. Among these the accuracy ratio, the capture rate, the hit rate, the gini index and the lift are the most popular ones (Gadi, Wang, & Lago, 2008; Kim & Han, 2003).

Parallel to its popularity, in the literature there are many studies on fraud detection using various data mining algorithms including decision trees, regression and artificial neural networks. Quah and Srinagesh (2008) suggest a framework which can be applied real time where first an outlier analysis is made separately for each customer using self organizing maps and then a predictive algorithm is utilized to classify the abnormal looking transactions. Panigrahi, Kundu, Sural, and Majumdar (2009) suggest a four component fraud detection solution which is connected in a serial manner. The main idea is first to determine a set of suspicious transactions and then run a Bayesian learning algorithm on this list to predict the frauds. Sanchez, Vila, Cerda, and Serrano (2009) presented a different approach and used association rule mining to define the patterns for normal card usage and indicating the ones not fitting to these patterns as suspicious. The study of Bolton and Hand (2002) provides a very good summary of literature on fraud detection problems. In these studies, the performance of the algorithms is mostly measured by the above measures.

When the fraudsters obtain a card, they usually use (spend) its entire available (unused) limit. According to the statistics, they do

* Corresponding author.
E-mail address: eduman@dogus.edu.tr (E. Duman).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.110

this in four or five transactions, on the average. Thus, for the fraud detection problem, although the above mentioned measures are quite relevant, as indicated by the bank authorities, a measure, measuring the loss that can be saved on the cards whose transactions are identified as fraud is more prominent. In other words, detecting a fraud on a card having a large available limit is more valuable than detecting a fraud on a card having a small available limit.

As a result, what we are faced with is a classification problem with variable misclassification costs. As the classical DM algorithms are not designed for such a misclassification cost structure, they are not directly applicable to our case (they work well when the objective is to minimize the incorrectly classified number of cases). Either some modifications should be made on them or new algorithms should be developed specifically for this purpose (actually in some popular DM software packages like SAS Enterprise Miner or SPSS PASW Modeler, it is possible to introduce different misclassification costs for the two classes but there has to a fixed ratio between them and thus they are not sufficient to handle our case).

As the classical DM algorithms are not directly usable, we need alternative methods for our classification problem. In this regard, we thought that, the meta-heuristic algorithms which are applicable to many different problem domains could serve. After analyzing the main characteristics of the popular meta-heuristic algorithms, for our problem we decided to use the genetic algorithm (GA) and the scatter search (SS) in a combined manner. We called our hybrid solution method as GASS.

Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses (Mitchell, 1998). Since their first introduction by Holland (1975), they have been successfully applied to many problem domains from astronomy (Charbonneau, 1995) to sports (Charbonneau, 1995), from optimization (Levi, Burrows, Fleming, & Hopkins, 2007; Krzysztof & Peter, 2004) to computer science (Kaya, 2010), etc. They have also been used in data mining mainly for variable selection (Bidgoli, Kashy, Kortemeyer, & Punch, 2003) and are mostly coupled with other DM algorithms.

Scatter search is another type of evolutionary algorithms. It has been first introduced by Glover (1977). Afterwards, it has been almost forgotten for about 20 years and since its re-introduction in 1997 (Glover, 1997) it has been applied to many different problems. However, to the best of our knowledge nobody has used it in DM problems so far.

The contributions of this study to the literature are twofold. First, a new classification cost function for the fraud detection problem is introduced. Secondly, a novel implementation of two well known meta-heuristic algorithms is made.

The rest of the paper is organized as follows. In the next section, the fraud detection problem we were faced is described in detail together with the current detection system used in our industrial partner. Section 3 briefly summarizes the basic principles of genetic algorithms and scatter search and then details the GASS implementation. The results obtained on the sample databases and the selections of the best solution parameters are discussed in Section 4. The sensitivity analysis regarding the parameter values is also made and presented in this section. The paper is finalized in Section 5 by providing the summary of the study and the major conclusions arrived.

2. Problem definition

There are two main types of CC frauds. First one is the counterfeit frauds which are carried out by organized criminal groups. Their total effect is huge and they are usually affecting tens and even hundreds of customers of a bank at a time. Until their next activity the fraudsters do usually remain inactive. The second type of CC fraud is the illegal use of a lost or stolen card. These type of frauds are mostly not related with criminal groups and each fraudster activity affects one or a few cards only.

The classical fraud detection solutions are expert rule systems based on rules that are produced by commonality and pattern detection analysis on previous fraud cases. However, local and transnational criminal groups are very dynamic in their structure and approaches. In this dynamic environment the power of even the best expert rules quickly deteriorate since the fraudsters start to behave different than the underlying pattern of the rule. In addition to this deficiency, these rules are only useful at the detection of counterfeit frauds but not useful at the detection of lost/stolen cases.

So what is needed indeed is a more robust solution which is not based only on the behavior of the fraudsters but on the behavior of customers also. Most customers have typical behaviors in using their cards and do not change their habits frequently. Thus, the typical behavior of each customer can be defined and any incoming transaction can be compared to that typical behavior. If it seems to be unusual it can be alerted as a possible fraud case.

An incoming transaction is subjected to the set of available rules and the points obtained from the rules satisfied are summed up to give the total suspiciousness points (TSP). If the TSP is greater than a predetermined threshold either the transaction is rejected or an alert is generated and the transaction is shown on the monitor of fraud monitoring staff who then decides what to do about that transaction. The possible actions are blocking the card, sending SMS or calling the merchant or the cardholder.

Our industrial partner developed its first version of fraud detection system in 2002 as a rule based system. In 2005 a new module was added which is based on customer behavioral analysis. The system has been in use since then. Some time ago we were asked to re-determine the levels and weights of parameters (variables) so as to minimize the loss due to frauds. Our scope was the improvement of the customer behavior related module of the existing solution. In this part many behavioral variables are evaluated and if the value of a variable is larger than the average plus a specified number of deviations (level), a certain suspiciousness point (weight) is generated. For the reasons of confidentiality we cannot give the list of variables here but we can say that they can be grouped into four: Those related to general (all) transactions statistics, regional statistics (statistics about the use in different geographical regions of the country), daily amount statistics (amount of transactions made in one day) and daily number statistics (number of transactions made in a day). The statistics about the MCC (merchant category code) and the country of the transactions were left out of the scope of the study. All together and including the threshold value for the TSP, we were expected to find the best values of 43 parameters (levels and weights).

The current values of parameters have been determined by some ad hoc procedures with the aim of maximizing the number of true alerts given that the number of alerts does not exceed a certain level (available monitoring capacity). If we define;

TP	the number of correctly classified alerts
TN	the number of correctly classified legitimates
FP	the number of transactions classified as fraudulent but are in fact legitimate
FN	the number of transactions classified as legitimate but are in fact fraudulent
r	the cost of monitoring an alert
TFL	the total amount of losses due to fraudulent transactions
S	savings in TFL with the use of fraud detection system

Then, TFL is the sum of the available limits of the cards whose transactions are labeled as TP or FN, and S = (available limits of

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات