New Perspectives of Pattern Recognition for Automatic Credit Card Fraud Detection

Addisson Salazar
Universitat Politècnica de València, Spain

Gonzalo Safont
Universitat Politècnica de València, Spain

Alberto Rodriguez
Universidad Miguel Hernández de Elche, Spain

Luis Vergara
Universitat Politècnica de València, Spain

INTRODUCTION

The automatic detection of frauds in financial operations using credit cards is a challenge issue that has been increasingly studied. The rapid expansion of information and communication technologies has expanded the potential to emulate legitimate operations by fraudsters. The solution to that problem has to be able to be adaptive since the behavior of frauds is changing constantly in time; to handle the detection in data with a very small ratio of fraud amount to legitimate operations, e.g., 5e-5; and accomplish operation requirements of very low false alarm ratios in real-time processing. Thus, several approaches have been proposed from pattern recognition and machine learning areas.

Main issues related with the problem of automatic credit card fraud detection (ACCFD) and proposed solutions are discussed from theoretical and practical standpoints. The perspective of detection analyses from receiving operating characteristic (ROC) curves and business key performance indicators (KPI) are jointly analyzed (Girgenti & Hedley, 2011) (Wells, 2011) (Montague, 2010). Therefore, a new conceptual framework for ACCFD considering modern techniques such as decision fusion and surrogate data is outlined. There are only a few references from the research field of signal processing for ACCFD, see for instance (Salazar, Safont, Soriano, & Vergara, 2012).

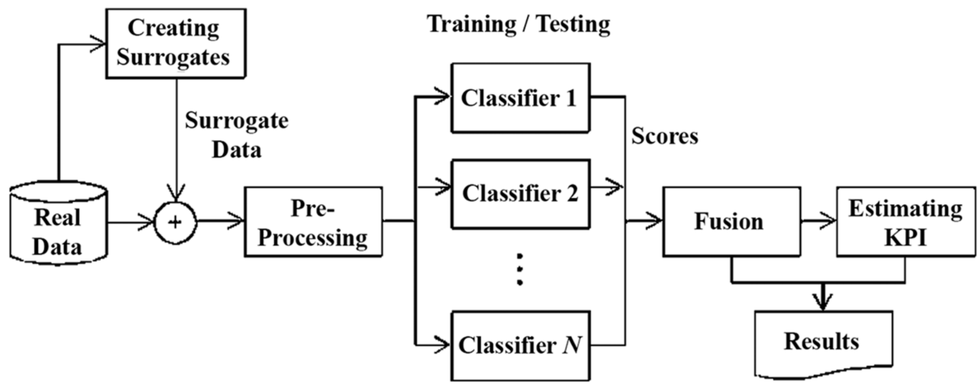
A case of study that combines different proportions of real and surrogate data is included. Several scenarios considering different single and combined methods are considered. ROC and KPI curves are analyzed bearing in mind numeric and operational requirements. The sensitivity of the methods to different proportions of fraud/legitimate ratios is tested. Thus, limitations and advantages of the studied methods are demonstrated.

BACKGROUND

Cyber-security and privacy have become very important subjects of research in recent years. This research spans many different fields, such as: security in the physical layer of wireless communications (Poor, 2012)); database security (Sankar, Rajagopalan, & Poor, 2013); distributed systems (Pawar, El Rouayheb, & Ramchandran, 2011); and biometrics (Lifeng, Ho, & Poor, 2011). One activity where the security and privacy mechanisms are critical is the e-commerce by using credit cards. This application features a massive volume of on-line transactions that are continuously exposed to frauds. Fraud detection

DOI: 10.4018/978-1-5225-2255-3.ch428

Figure 1. Outline of the signal processing procedure



in credit card transactions is a critical problem affecting large financial companies and involving annually loss of billions of dollars (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011).

Basically two strategies can be raised. The first consists of defining the problem as one-class classification, and thus, characterizing the largest data population (the legitimate transactions) and considering all the data with different characteristics as outliers (Hodge & Austin, 2004) (Tax & Duin, 2001). The second strategy is to define the problem as a two-class classification characterizing legitimate and fraudulent transaction data. We have concentrated in this later detection approach which takes full advantage of the available labeled data.

There is extensive literature that reviews and provides taxonomies and comparisons about the large number of ACCFD methods that have been developed during the last two decades (e.g., (Danenas, 2015)). However, only few of these references are from the research field of signal processing. The particular characteristics of ACCFD make this a challenging problem for signal processing algorithms (Salazar, Safont, Soriano, & Vergara, 2012). Optimum design of the algorithms depends on the detection models employed to estimate the multidimensional joint distribution of the random variables underlying the data.

Figure 1 shows an outline of the proposed signal processing procedure. The multivariate surrogate data is obtained following the methods explained in (Salazar, Safont, & Vergara, Surrogate techniques for testing fraud detection algorithms in credit card operations, 2014). The pre-processing step consists of applying principal component analysis (PCA) to reduce dimensionality of the data preserving 95% of data variance.

SURROGATE DATA

Bank enterprises collect large amount of historical records corresponding to millions of credit cards operations, but, unfortunately, only a small portion, if any, is open access. This is because, e.g., the records include confidential customer data and banks are afraid of public quantitative evidence of existing fraud operations (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011). A solution is to generate synthetic records which replicate as much as possible the behavior of the real data. Surrogate techniques give an approach to this problem. Surrogates algorithms have been extensively used to detect the possible presence of non-linearities in a given time series realization. Basically, surrogate replicas of the original data are generated trying to preserve the correlation (second-order statistic) and amplitude distribution (first-order