# Abstract

With increasing adaptation of electronic payment and e-commerce industry, fraud detection has becomes an increasing critical factor.

In this paper, we

The actions taken against fraud can be divided into:

* Fraud prevention, which attempts to block fraudulent transactions at source, and fraud detection, where successful fraud transactions are identified a posteriori.
  + Address Verification Systems (AVS), Card Verification Method (CVM)  
    and Personal Identification Number (PIN).
* Fraud detection is, given a set of credit card transactions, the process of identifying if  
  a new authorized transaction belongs to the class of fraudulent or genuine transactions .[4]
* A Fraud Detection System (FDS) should not only detect fraud cases efficiently.

With Machine Learning (ML) techniques [7] we can efficiently  
discover fraudulent patterns and predict transactions that are most likely to be fraudulent. ML techniques consist in inferring a prediction model on the basis of a set of  
examples. The model is in most cases a parametric function, which allows predicting the  
likelihood of a transaction to be fraud, given a set of features describing the transaction.  
In the domain of fraud detection, the use of learning techniques is attractive for a number of reasons. First, they allow to discovery patterns in high dimensional data streams,  
i.e. transactions arrive as a continuous stream and each transaction is defined by many  
variables. Second, fraudulent transactions are often correlated both over time and space.  
For examples, fraudsters typically try to commit frauds in the same shop with different  
cards within a short time period. Third, learning techniques can be used to detect and  
model existing fraudulent strategies as well as identify new strategies associated to unusual behaviour of the cardholders. Predictive models based on ML techniques are also  
able to automatically integrate investigators’ feedbacks to improve the accuracy of the  
detection, while in the case of expert system, including investigators feedbacks requires  
rules revision that can be tedious and time consuming.

Though, this field has been studied from some decades now with many purposals being put forward, still fraud detection in credit card transaction remains a relevant problem today. This is mainly due to the fact that many learning algorithm relies on assumptions that hardly hold in real-world Fraud Detection System (FDS). This lack of realism concerns mainly following aspects:

1. Genuine Transactions far outnumber frauds.
2. Transactions might change their statistical properties over time.
3. The way and timing with which supervised information is provided by the expert investigators.

This paper tries to model the real-world FDS based on the proposed model.

Understanding the sampling methods

Form

after the proposed model in[9], This realistically describes the operating conditions of FDSs that analyze massive streams of e-trasactions.

Based on the model mentioned above, we design and asses a learning strategy [10]which effectively addresses the problems mentioned for Real-time FDS, namely, concept drift, verification latency and class imbalance.

The strategy is run on the credit card transaction data that in available in pubic domain[11].

We present a prototype of the FDS able to meet real-work working condition.

Fraud detection differs from conventional classification tasks because in a first phase only a small set of supervised samples is provided by human investigators.

Labels of the vast majority of transactions are made available only several days later, when customers  
have possibly reported unauthorized transactions. The delay in obtaining accurate labels and the interaction between alerts and supervised information has to be carefully  
taken into consideration when learning in a concept-drifting environment. We show that  
investigator’s feedbacks and delayed labels have to be handled separately and require  
different classification strategies. Finally, based on our results, we argue that the best  
detection solution consists in training two separate classifiers (on feedbacks and delayed  
labels, respectively), and then aggregating the outcomes.

Software and Credit Card Fraud Detection Dataset.

1. For classification problems we use a software package called unbalanced[24] shared in the paper titled “” available in R language. It implements some techniques for unbalanced classification tasks and provides a racing strategy [227] to adaptively select the best methods for a given dataset, for a given dataset, classification algorithms and accuracy measure adopted.
2. For using dataset containing the information about credit card transactions with examples of fraudulent samples. The  
   dataset was first used in the experiments of our paper [28] and then made available  
   at: http://www.ulb.ac.be/di/map/adalpozz/data/creditcard.Rdata. It contains 31  
   numerical input variables. The dataset is highly unbalanced; frauds represent 0*:*172% of all transactions (492 frauds  
   out of 284807 transactions). This is one of the rare datasets on fraud detection available to the community.

**2.2.1 Fraud Detection System working conditions**

*Data Driven Model (DDM)* layer relies on a predictive model to assign a fraud score to  
each transaction. Usually this phase uses ML algorithms that return for each transaction an estimate of the probability to be a fraud. These algorithms can learn complex  
correlations in the data using large volume of data with high dimensionality. They are usually more robust than EDR, but most of them are black box, i.e. it is not possible  
to convert them into rules that are easy to interpret. DDMs are able to consider all  
the information associated to a transaction, while EDR returns conditions based on few  
features of the transactions. The focus of the thesis is to improve the DDM part of the FDS and to model the  
interaction between Data Driven Methods based on ML and investigators. In particular,  
we will consider a scenario where the DDM is the only element of the FDS responsible for  
the alerts given to fraud experts and the algorithms are able to learn from the feedback  
provided.

Most FDSs monitor streams of credit card transactions by means of classifiers returning  
alerts for the riskiest payments. Fraud detection differs from conventional classification  
because, in a first phase, human investigators who have time to assess only a reduced  
number of alerts providing a small set of supervised samples denoted as feedbacks. Labels  
of the vast majority of transactions are made available only several days later, when  
customers have possibly reported unauthorized transactions. These transactions define  
an additional set of supervised samples called delayed samples. The delay in obtaining  
accurate labels and the interaction between alerts and supervised information has to be  
carefully taken into consideration when learning in a concept-drifting environment.  
In this chapter we address a realistic fraud detection setting and we show that feedbacks  
and delayed samples have to be handled separately. We design two prototypes of FDS  
on the basis of an ensemble and a sliding-window approach (Section 6.3.1) and we show  
that the winning strategy consists in training two separate classifiers (on feedbacks and  
delayed samples, respectively), and then aggregating the outcomes. Experiments on large datasets show that the alert precision, which is the primary concern of investigators, can  
substantially be improved by the proposed approach.  
In order to obtain precise alerts, feedbacks samples have to receive larger weights than  
non-alerted transactions and methods that diminish their role in the learning process  
lead to loss of predictive accuracy

# Introduction

The expansion of the electronic commerce, together with increasing confidence of customers in electronic payment, makes fraud detection a critical factor. Detection frauds in real time setting demands the design and implementation of scalable learning techniques able to ingest and analyse massive amounts of streaming data.

The need of automatic systems to able to detect frauds from historical data led to design of number of machine learning algorithms for fraud detection. These automatic systems are essential since it is not always possible or easy for human analyst to detect fraudulent patterns in transactional dataset, often characterized by a large number of samples, many dimensions and online updates.

In order to minimize cost of detection it is important to use some expert rules and statistical based models (e.g. Machine Learning) to make first screen between genuine and potential fraud and ask investigators to review only the cases with high risk.

This feature is provided by the predictive model. The predictive model scores each transactions with high or low risk fraud and those with high risk generate alerts. Investigators check these alerts and provide feedback on each alert, ie. True positive (fraud) or false positive (genuine). These feedback can then be used to improve the model.

The Machine Learning techniques applied in this predictive model can also efficiently discover fraudulent patterns and predict transactions that are most likely to be fraudulent. ML techniques consists in inferring a prediction model on the basis of set of examples.

In the domain of fraud detection, the use of learning techniques is attractive for a number of reasons.

* First, they allow to discovery patterns in high dimensional data streams, i.e. transactions arrive as a continuous stream and each transaction is defined by many variables.
* Second, fraudulent transactions are often correlated both over time and space. For examples, fraudsters typically try to commit frauds in the same shop with different cards within a short time period.
* Third, learning techniques can be used to detect and model existing fraudulent strategies as well as identify new strategies associated to unusual behaviour of the cardholders.

Predictive models based on ML techniques are also able to automatically integrate investigators’ feedbacks to improve the accuracy of the detection, while in the case of expert system, including investigators feedbacks requires rules revision that can be tedious and time consuming.

# Objective

Objective of this submission is to model the realistic Fraud Detection System (FDS) based on previous research paper. Based on this realisitic FDS, we suggest the Preditive Model that implements the the Supervised Machine Learning Techniques to solve this Real FDS Problem.

# Innovation

Implementation of the better Predictive Model for Fraud Detection System(FDS) based on Supervised Machine Learning Techniques. This Predictive model based on supervised ML technique is able to automatically integrate investigators’ feedbacks to improve the accuracy of the detection.

We demonstrate this on the data available on the public domain regarding credit card transactions.

# Overview

# Summary/ Results:

In Progress.

Developing the Supervised machine Learning Predictive Model to be run on the data.