Learned lessons in credit card fraud detection from a practitioner perspective

Andrea Dal Pozzolo [a,*], Olivier Caelen [b], Yann-Aël Le Borgne [a], Serge Waterschoot [b], Gianluca Bontempi [a]

[a] *Machine Learning Group, Computer Science Department, Faculty of Sciences ULB, Université Libre de Bruxelles, Brussels, Belgium*
[b] *Fraud Risk Management Analytics, Worldline, Brussels, Belgium*

ARTICLE INFO

ABSTRACT

Billions of dollars of loss are caused every year due to fraudulent credit card transactions. The design of efficient fraud detection algorithms is key for reducing these losses, and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to non-stationary distribution of the data, highly imbalanced classes distributions and continuous streams of transactions.

At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about which is the best strategy to deal with them.

In this paper we provide some answers from the practitioner's perspective by focusing on three crucial issues: unbalancedness, non-stationarity and assessment. The analysis is made possible by a real credit card dataset provided by our industrial partner.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays, enterprises and public institutions have to face a growing presence of fraud initiatives and need automatic systems to implement fraud detection (Delamaire, Abdou, & Pointon, 2009). Automatic systems are essential since it is not always possible or easy for a human analyst to detect fraudulent patterns in transaction datasets, often characterized by a large number of samples, many dimensions and online updates. Also, the cardholder is not reliable in reporting the theft, loss or fraudulent use of a card (Pavía, Veres-Ferrer, & Foix-Escura, 2012). Since the number of fraudulent transactions is much smaller than the legitimate ones, the data distribution is unbalanced, i.e. skewed towards non-fraudulent observations. It is well known that many learning algorithms underperform when used for unbalanced dataset (Japkowicz & Stephen, 2002) and methods (e.g. resampling) have been proposed to improve their performances. Unbalancedness is not the only factor that determines the difficulty of a classification/detection task. Another influential factor is the amount of overlapping of the classes of interest due to limited information that transaction records provide about the nature of the process (Holte, Acker, & Porter, 1989).

Detection problems are typically addressed in two different ways. In the static learning setting, a detection model is periodically relearnt from scratch (e.g. once a year or month). In the online learning setting, the detection model is updated as soon as new data arrives. Though this strategy is the most adequate to deal with issues of non stationarity (e.g. due to the evolution of the spending behavior of the regular card holder or the fraudster), little attention has been devoted in the literature to the unbalanced problem in changing environment.

Another problematic issue in credit card detection is the scarcity of available data due to confidentiality issues that give little chance to the community to share real datasets and assess existing techniques.

2. Contributions

This paper aims at making an experimental comparison of several state of the art algorithms and modeling techniques on one real dataset, focusing in particular on some open questions like: Which machine learning algorithm should be used? Is it enough to learn a model once a month or it is necessary to update the model everyday? How many transactions are sufficient to train the model? Should the data be analyzed in their original unbalanced form? If not, which is the best way to rebalance them? Which performance measure is the most adequate to asses results?

In this paper we address these questions with the aim of assessing their importance on real data and from a practitioner

* Corresponding author. Tel.: +32 2 650 55 94.
E-mail addresses: adalpozz@ulb.ac.be (A. Dal Pozzolo), olivier.caelen@worldline.com (O. Caelen), yleborgn@ulb.ac.be (Y.-A. Le Borgne), serge.waterschoot@worldline.com (S. Waterschoot), gbonte@ulb.ac.be (G. Bontempi).