Expert Systems With Applications 51 (2016) 134–142

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature engineering strategies for credit card fraud detection

Alejandro Correa Bahnsen*, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg

ARTICLE INFO

Keywords:
Cost-sensitive learning
Fraud detection
Preprocessing
Von Mises distribution

ABSTRACT

Every year billions of Euros are lost worldwide due to credit card fraud. Thus, forcing financial institutions to continuously improve their fraud detection systems. In recent years, several studies have proposed the use of machine learning and data mining techniques to address this problem. However, most studies used some sort of misclassification measure to evaluate the different solutions, and do not take into account the actual financial costs associated with the fraud detection process. Moreover, when constructing a credit card fraud detection model, it is very important how to extract the right features from the transactional data. This is usually done by aggregating the transactions in order to observe the spending behavioral patterns of the customers. In this paper we expand the transaction aggregation strategy, and propose to create a new set of features based on analyzing the periodic behavior of the time of a transaction using the von Mises distribution. Then, using a real credit card fraud dataset provided by a large European card processing company, we compare state-of-the-art credit card fraud detection models, and evaluate how the different sets of features have an impact on the results. By including the proposed periodic features into the methods, the results show an average increase in savings of 13%.

1. Introduction

The use of credit and debit cards has increased significantly in the last years, unfortunately so has fraud. Because of that, billions of Euros are lost every year. According to the European Central Bank (European Central Bank, 2014), during 2012 the total level of fraud reached 1.33 billion Euros in the Single Euro Payments Area, which represents an increase of 14.8% compared with 2011. Moreover, payments across non traditional channels (mobile, internet, etc.) accounted for 60% of the fraud, whereas it was 46% in 2008. This opens new challenges as new fraud patterns emerge, and current fraud detection systems are less successful in preventing these frauds.

Furthermore, fraudsters constantly change their strategies to avoid being detected, something that makes traditional fraud detection tools such as expert rules inadequate (Van Vlasselaer et al., 2015), moreover, machine learning methods as well can be inadequate if they miss to adapt to new fraud strategies, i.e., static models that are never updated (Dal Pozzolo, Caelen, Le Borgne, Waterschoot, & Bontempi, 2014).

The use of machine learning in fraud detection has been an interesting topic in recent years. Several detection systems based on machine learning techniques have been successfully used for this problem (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011). When constructing a credit card fraud detection model, there are several factors that have an important impact during the training phase: Skewness of the data , cost-sensitivity of the application, short-time response of the system, dimensionality of the search space and how to preprocess the features (Bachmayer, 2008; Bolton, Hand, Provost, & Breiman, 2002; Dal Pozzolo et al., 2014; Van Vlasselaer et al., 2015; Whitrow, Hand, Juszczak, Weston, & Adams, 2008). In this paper, we address the cost-sensitivity and the features preprocessing to achieve improved fraud detection and savings.

Credit card fraud detection is by definition a cost-sensitive problem, in the sense that the cost due to a false positive is different than the cost of a false negative. When predicting a transaction as fraudulent, when in fact it is not a fraud, there is an administrative cost that is incurred by the financial institution. On the other hand, when failing to detect a fraud, the amount of that transaction is lost (Hand, Whitrow, Adams, Juszczak, & Weston, 2007). Moreover, it is not enough to assume a constant cost difference between false positives and false negatives, as the amount of the transactions varies quite significantly; therefore, its financial impact is not constant but depends on each transaction. In Correa Bahnsen, Stojanovic, Aouada, and Ottersten (2013), we proposed a new cost-based measure to evaluate credit card fraud detection models, taking into account the different financial costs incurred by the fraud detection process.

* Corresponding author. Tel.: +57 3045462842.
E-mail addresses: al.bahnsen@gmail.com (A. Correa Bahnsen), djamila.aouada@uni.lu (D. Aouada), aleksandar.stojanovic@rwth-aachen.de (A. Stojanovic), bjorn.ottersten@uni.lu (B. Ottersten).

http://dx.doi.org/10.1016/j.eswa.2015.12.030
0957-4174/

When constructing a credit card fraud detection model, it is very important to use those features that allow accurate classification. Typical models only use raw transactional features, such as time, amount, place of the transaction. However, these approaches do not take into account the spending behavior of the customer, which is expected to help discover fraud patterns (Bachmayer, 2008). A standard way to include these behavioral spending patterns is proposed in (Whitrow et al., 2008), where Whitrow et al. proposed a transaction aggregation strategy in order to take into account a customer spending behavior. The computation of the aggregated features consists in grouping the transactions made during the last given number of hours, first by card or account number, then by transaction type, merchant group, country or other, followed by calculating the number of transactions or the total amount spent on those transactions.

In this paper we first propose a new savings measure based on comparing the financial cost of an algorithm versus using no model at all. Then, we propose an expanded version of the transaction aggregation strategy, by incorporating a combination criteria when grouping transactions, i.e., instead of aggregating only by card holder and transaction type, we combine it with country or merchant group. This allows to have a much richer feature space.

Moreover, we also propose a new method for extracting periodic features in order to estimate if the time of a new transaction is within the confidence interval of the previous transaction times. The motivation is that a customer is expected to make transactions at similar hours. The proposed methodology is based on analyzing the periodic behavior of a transaction time, using the von Mises distribution (Fisher, 1995).

Furthermore, using a real credit card fraud dataset provided by a large European card processing company, we compare the different sets of features (raw, aggregated, extended aggregated and periodic), using two kind of classification algorithms; cost-insensitive (Hastie, Tibshirani, & Friedman, 2009) and example-dependent cost-sensitive (Elkan, 2001). The results show an average increase in the savings of 13% by using the proposed periodic features. Additionally, the outcome of this paper is being currently used to implement a state-of-the-art fraud detection system, that will help to combat fraud once the implementation stage is finished.

The remainder of the paper is organized as follows. In Section 2, we explain the background on credit card fraud detection, and specifically the measures to evaluate a fraud detection model. Then in Section 3, we discuss current approaches to create the features used in fraud detection models, moreover, we present our proposed methodology to create periodic based features. Afterwards, the experimental setup is given in Section 4. In Section 5, the results are shown. Finally, conclusions and discussions of the paper are presented in Section 6.

2. Credit card fraud detection evaluation

A credit card fraud detection algorithm consists in identifying those transactions with a high probability of being fraud, based on historical fraud patterns. The use of machine learning in fraud detection has been an interesting topic in recent years. Different detection systems that are based on machine learning techniques have been successfully used for this problem, in particular: neural networks (Maes, Tuyls, Vanschoenwinkel, & Manderick, 2002), Bayesian learning (Maes et al., 2002), artificial immune systems (Bachmayer, 2008), association rules (Sánchez, Vila, Cerda, & Serrano, 2009), hybrid models (Krivko, 2010), support vector machines (Bhattacharyya et al., 2011), peer group analysis (Weston, Hand, Adams, Whitrow, & Juszczak, 2008), random forest (Correa Bahnsen et al., 2013; Dal Pozzolo et al., 2014), discriminant analysis (Mahmoudi & Duman, 2015) and social network analysis (Van Vlasselaer et al., 2015).

Table 1
Classification confusion matrix.

	Actual positive $y = 1$	Actual negative $y = 0$
Predicted positive $c = 1$	True positive (*TP*)	False positive (*FP*)
Predicted negative $c = 0$	False negative (*FN*)	True positive (*TN*)

Table 2
Cost matrix (Elkan, 2001).

	Actual positive $y_i = 1$	Actual negative $y_i = 0$
Predicted positive $c_i = 1$	C_{TP_i}	C_{FP_i}
Predicted negative $c_i = 0$	C_{FN_i}	C_{TN_i}

Most of these studies compare their proposed algorithms with a benchmark algorithm and then make the comparison using a standard binary classification measure, such as misclassification error, receiver operating characteristic (*ROC*), Kolmogorov–Smirnov (*KS*), $F_1 Score$ (Bolton et al., 2002; Hand et al., 2007) or AUC statistics (Dal Pozzolo et al., 2014). Most of these measures are extracted by using a confusion matrix as shown in Table 1, where the prediction of the algorithm c_i is a function of the k features of transaction i, $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^k]$ and y_i is the true class of the transaction i.

From this table, several statistics are extracted. In particular:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Recall = \frac{TP}{TP+FN}$
- $Precision = \frac{TP}{TP+FP}$
- $F_1 Score = 2\frac{Precision \cdot Recall}{Precision+Recall}$

However, these measures may not be the most appropriate evaluation criteria when evaluating fraud detection models, because they tacitly assume that misclassification errors carry the same cost, similarly with the correct classified transactions. This assumption does not hold in practice, when wrongly predicting a fraudulent transaction as legitimate carries a significantly different financial cost than the inverse case. Furthermore, the accuracy measure also assumes that the class distribution among transactions is constant and balanced (Provost, Fawcett, & Kohavi, 1998), and typically the distributions of a fraud detection dataset are skewed, with a percentage of frauds ranging from 0.005% to 0.5% (Bachmayer, 2008; Bhattacharyya et al., 2011).

In order to take into account the different costs of fraud detection during the evaluation of an algorithm, we may use the modified cost matrix defined in (Elkan, 2001). In Table 2, the cost matrix is presented, where the cost assof correct classification, namely, true positives C_{TP_i}, and true negatives C_{TN_i}; and the two types of misclassification errors, namely, false positives C_{FP_i}, and false negatives C_{FN_i}, are presented. This is an extension of Table 1, but in this case the costs are example-dependent, in other words, specific to each transaction i.

Hand et al. (Hand et al., 2007) proposed a cost matrix, where in the case of false positive the associated cost is the administrative cost $C_{FP_i} = C_a$ related to analyzing the transaction and contacting the card holder. This cost is the same assigned to a true positive $C_{TP_i} = C_a$, because in this case, the card holder will have to be contacted. However, in the case of a false negative, in which a fraud is not detected, the cost is defined to be a hundred times larger,