



Multiple perspectives HMM-based feature engineering for credit card fraud detection

Yvan Lucas
INSA Lyon
Universität Passau
yvanlucas44@gmail.com

Pierre-Edouard Portier
INSA Lyon
pierre-edouard.portier@insa-lyon.fr

Léa Laporte
INSA Lyon
lea.laporte@insa-lyon.fr

Sylvie Calabretto
INSA Lyon
sylvie.calabretto@insa-lyon.fr

Olivier Caelen
ATOS Worldline
o.caelen@gmail.com

Liyun He-Guelton
ATOS Worldline
liyun.he-guelton@worldline.com

Michael Granitzer
Universität Passau
michael.granitzer@uni-passau.de

ABSTRACT

Machine learning and data mining techniques have been used extensively in order to detect credit card frauds. However, most studies consider credit card transactions as isolated events and not as a sequence of transactions.

In this article, we model a sequence of credit card transactions from three different perspectives, namely (i) does the sequence contain a Fraud? (ii) Is the sequence obtained by fixing the card-holder or the payment terminal? (iii) Is it a sequence of spent amount or of elapsed time between the current and previous transactions? Combinations of the three binary perspectives give eight sets of sequences from the (training) set of transactions. Each one of these sets is modelled with a Hidden Markov Model (HMM). Each HMM associates a likelihood to a transaction given its sequence of previous transactions. These likelihoods are used as additional features in a Random Forest classifier for fraud detection. This multiple perspectives HMM-based approach enables an automatic feature engineering in order to model the sequential properties of the dataset with respect to the classification task. This strategy allows for a 15% increase in the precision-recall AUC compared to the state of the art feature engineering strategy for credit card fraud detection.

KEYWORDS

Machine Learning, Credit Card Fraud Detection, Hidden Markov Models, Random Forest, Sequence classification

ACM Reference Format:

Yvan Lucas, Pierre-Edouard Portier, Léa Laporte, Sylvie Calabretto, Olivier Caelen, Liyun He-Guelton, and Michael Granitzer. 2019. Multiple perspectives HMM-based feature engineering for credit card fraud detection. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19)*, April 8–12, 2019, Limassol, Cyprus. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3297280.3297586>

1 INTRODUCTION

Credit card fraud detection presents several difficulties. One of them is the fact that the feature set describing a credit card transaction usually ignores detailed sequential information. Typical models only use raw transactional features, such as time, amount, merchant category, etc. Bolton & al. [2] showed the necessity to use attributes describing the history of the transaction when they used unsupervised methods such as peer group analysis for credit card fraud detection. Consequently, Whitrow & al. [4] create descriptive statistics as features in order to include historical knowledge. These descriptive features can be for example the number of transactions or the total amount spent from the card-holder in the past 24 hours with the same merchant category or country. Among other authors, Bahnsen & al. [1] established Whitrow's transaction aggregation strategy as the state of the art feature engineering technique for credit card fraud detection.

We identified several weaknesses in the construction of these features that motivated our work: Descriptive statistics provide an aggregated view over a set of transactions. Such aggregated features do not consider fine-grained temporal dependencies between the transactions. For example, a common fraud pattern starts with low amount transactions for testing the card, followed by high amount transaction. Moreover, these aggregated features consider only the history of the card-holder and do not exploit information of fraudulent transactions for feature engineering. However, a sequence of transactions happening at a fixed terminal can also contain valuable patterns for fraud detection.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SAC '19, April 8–12, 2019, Limassol, Cyprus
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5933-7/19/04.
<https://doi.org/10.1145/3297280.3297586>

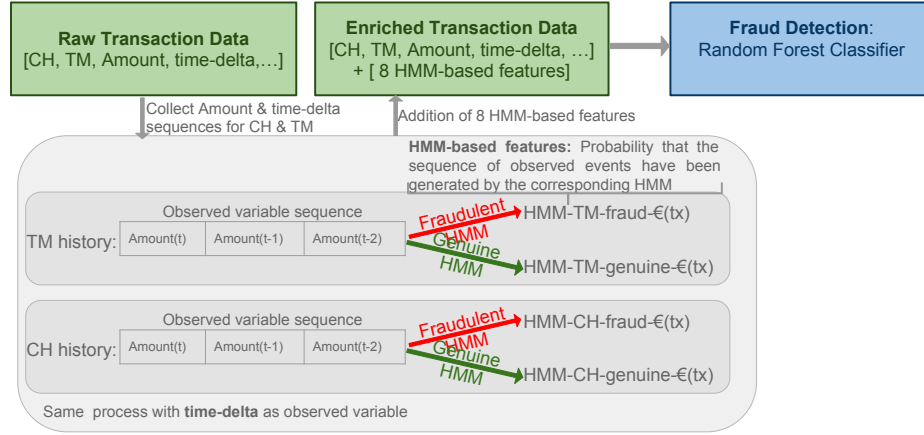


Figure 1: Enriching transaction data with HMM-based features calculated from multiple perspectives (*CH*=Card-holder, *TM*=Terminal)

In our work we propose to generate history-based features using Hidden Markov models. These features are created by estimating the likelihood of a sequence of transactions to be regular in regard to terminal or cardholder transactions. More precisely, they quantify the similarity between an observed sequence and the sequence of past fraudulent or genuine transactions observed for the cardholders or the terminals.

2 MULTIPLE PERSPECTIVE HMM-BASED FEATURE ENGINEERING

The state of the art feature engineering techniques for credit card fraud detection creates descriptive features using the history of the card-holder (such as: "amount spent by the card-holder in shops from a given country in the last 24h" [4] [1]). These descriptive features present several limits we aim to overcome. First they do not take into account the history of the seller even if it is clearly identified in most credit card transactions dataset. Moreover these descriptive features do not consider dependencies between transactions of a same sequence. Therefore we use Hidden Markov Models which are generative probabilistic models and a common choice for sequence modelling [?].

In addition to the descriptive aggregated features created by Whitrow & al. [4], we propose to create eight new HMM-based features. They quantify the similarity between the history of a transaction and eight distributions learned previously on set of sequences selected in a supervised way in order to model different perspectives.

In particular, we select three perspectives for modelling a sequence of transactions (see figure 1). A sequence (i) can be made only of genuine historical transactions or can include at least one fraudulent transaction in the history, (ii) can come from a fixed card-holder or from a fixed terminal, and (iii) can consist of amount values or of time-delta values (i.e. the difference in time between the current transaction and the previous one). We optimised the parameters of eight HMMs

using all eight possible combinations (i-iii). The HMM-based features proposed in this paper are the likelihoods that a sequence is generated by each of these models.

In order to make the HMMs model the genuineness and fraudulence of the card holders and the terminals, we create 4 training set containing:

- (1) Sequences of transactions from **genuine credit cards** (without fraudulent transactions in their history).
- (2) Sequences of transactions from **genuine terminals** (without fraudulent transactions in their history)
- (3) Sequences of transactions from **compromised credit cards** (with at least one fraudulent transaction)
- (4) Sequences of transactions from **compromised terminals** (with at least one fraudulent transaction)

We then extract from these sequences of transactions the symbols that will be the observed variable for the HMMs. In our experiments, the observed variable can be either:

- (1) The amount of a transaction.
- (2) The amount of time elapsed between two consecutive transactions of a card-holder (time-delta).

At the end, we obtain 8 trained HMMs modeling 4 types of behaviour (genuine terminal behaviour, fraudulent terminal behaviour, genuine card-holder behaviour and fraudulent card-holder behaviour) for both observed variables (amount and time-delta).

The HMM-based features are the likelihood that the recent sequence of observed events has been generated by a given HMM.

3 EXPERIMENTAL SETUP

We use the Python library *hmmlearn*¹.

The value of each HMM-based feature is the likelihood that the sequence made of the current transaction and the two previous ones from this terminal/card holder has been generated by the corresponding HMM.

¹<https://github.com/hmmlearn/hmmlearn>

In order for the HMM-based features and the aggregated features to be comparable, we calculate terminal-centered aggregated features in addition to the card-holder centered aggregated features (see table 1)

Feature	Signification
AGGCH1	# transactions issued by user in 24h.
AGGCH2	Amount spent by user in 24h.
AGGCH3	# transactions in the country in 24h.
AGGCH4	Amount spent in the country in 24h.
AGGTM1	# transactions in terminal in 24h.
AGGTM2	Amount spent in terminal in 24h.
AGGTM3	# transactions with this card type in 24h.
AGGTM4	Amount spent with this card type in 24h.

Table 1: Aggregated features centered on the card holders and the terminal

We used a credit card transactions dataset provided by our industrial partner in order to quantify the increase in detection when adding HMM-based features. This dataset contains $4.7 * 10^7$ anonymized transactions from the belgian credit cards between 01.03.2015 and 31.05.2015. We split temporally the dataset in three different parts: the training set, the validation set and the testing set.

We tune the Random Forest hyperparameters through a grid search that optimizes the Precision-Recall Area under the Curve on the validation set. The choice of the Precision-Recall AUC for imbalanced dataset was motivated by the work of Davis & al. [3].

4 IMPROVEMENT IN FRAUD DETECTION WHEN USING HMM-BASED FEATURES

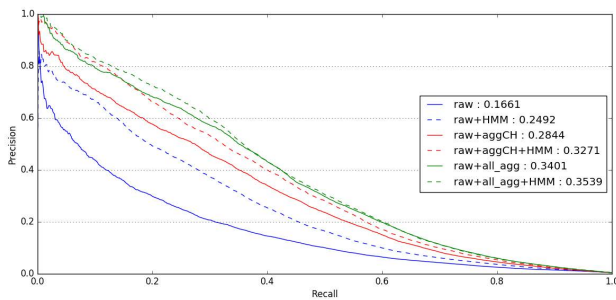


Figure 2: Predictions using HMM-based features

We train Random Forest Classifiers using different feature sets in order to compare the efficiency of prediction when we add HMM-based features to the classification task.

We tested the addition of our HMM-based features to several feature sets. We refer to the feature set "raw+aggCH" as the state of the art feature engineering strategy since it contains all the raw features with the addition of Whitrow's

aggregated features [4]. The feature groups we refer to are: the raw features (raw), the features based on the aggregations of card-holders transactions (aggCH), the features based on the aggregation of terminal transactions (aggTM), the proposed HMM-based features (HMM features).

In this section, the HMMs were created with 5 hidden states and the HMM-based features were calculated with a window-size of 3 (actual transaction + 2 past transactions of the card-holder and of the terminal).

By comparing the AUC of the curves raw+aggCH and raw+aggCH+HMM, we observe that adding HMM-based features to the state of the art feature engineering strategy introduced in the work of Whitrow & al. [4] leads to an increase of 15.1% of the PR-AUC.

The addition of features that describe the sequence of transactions, be it HMM-based features or Whitrow's aggregated features, increases a lot the detection.

5 CONCLUSION

The multiple perspective property of our HMM-based feature engineering strategy gives us the possibility to incorporate a broad spectrum of sequential information. In fact, we model the genuine and fraudulent behaviours of the merchants and the card-holders according to two features: the timing and the amount of the transactions. Moreover, the HMM-based features are created in a supervised way and therefore lower the need of expert knowledge for the creation of the fraud detection system.

The results show an increase in the precision-recall AUC of 15.1% due to the addition of our multi-perspective HMM-based features when compared to the state of the art feature engineering strategies.

HMM-based feature engineering strategy is a powerful tool that is shown to present interesting properties for fraud detection. We can imagine building similar HMM-based features in any supervised task that involve a sequential dataset.

To ensure reproducibility, an optimized code for calculating and evaluating the proposed HMM-based features can be found at https://gitlab.com/Yvan_Lucas/hmm-ccfd.

5.1 Acknowledgement:

The work has been funded partially by the Bavarian Ministry of Economic Affairs, Regional Development and Energy in the project Internetkompetenzzentrum Ostbayern.

REFERENCES

- [1] Bahnsen A. C., Aouada D., Stojanovic A., and Ottersten B. (2016) Feature engineering strategies for credit card fraud detection. *Expert Systems With Applications*.
- [2] Bolton R. and Hand D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*.
- [3] Davis J. and Goadrich M. (2006). The relationship between precision-recall and roc curves. *ICML 06 Proceedings of the 23rd international conference on Machine learning*.
- [4] Whitrow C., Hand D. J., Juszczak P., Weston D. J., and Adams N. M. (2008). Transaction aggregation strategy for credit card fraud detection. *Data Mining and Knowledge Discovery* 18(1).