Transactions Fraud Detection using Machine Learning and Nature Inspired Algorithms

Peter Mačinec and Timotej Zaťko

Faculty of Informatics and Information Technologies,
Slovak University of Technology, Bratislava

Abstract. The abstract should briefly summarize the contents of the paper in 15–250 words.

Keywords: Transactions Fraud Detection · Machine Learning · Nature Inspired Algorithms · Data Analysis.

1 Introduction

Payments with credit cards are nowadays still more preferred over using cash. Using credit card is not only more comfortable for people, but even also more safe than carrying cash in wallet when it comes to higher amount of money. However, number of transaction frauds is arising alongside with usage of credit cards. Number of transactions and ability to obtain data from them indicate the need of automatic detection of suspicious payments.

Data of performed transactions via credit cards are naturally collected by companies and banks to produce statistics or investigate frauds. Therefore, transactions fraud detection can be interpreted as data mining problem, concretely binary classification.

In this paper, we propose novel method for detection frauds in transactions using aspects of data analysis, machine learning and nature inspired algorithms. The basis of our method lies in training machine learning model on best features selected by nature inspired algorithm. More nature inspired algorithms are compared to choose the best one for this problem. Finally, nature inspired algorithms in combination with machine learning proved to be very efficient way for detecting frauds in transactions, overperforming common methods in this area.

2 Related works

- Gray Wolf Optimization Algorithm [1]

3 Problem definition

Majority of datasets available for data science research of transactions fraud detection have the same characteristics - highly imbalanced data, a lot of features and majority of features are anonymized.

Our method will be trained and evaluated on dataset from Kaggle competition EEE-CIS Fraud Detection[1]. As usual in problems of transactions fraud detection, a lot of features of different types are available - 434 features describing demography, credit card, transaction itself, etc. Majority of features are anonymized or the meaning of them is not clear. In this case, we must be careful to avoid labels leak from some features and also in interpreting data analysis or models result when talking about anonymized features. This dataset contains almost 600k samples, that is appropriate for machine learning algorithms. However, data are highly imbalanced - classes distribution can bee seen at figure 1. When training and evaluating models, one should be careful when data are imbalanced.

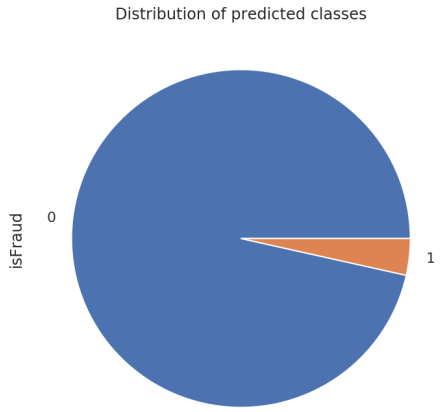

Fig. 1. Predicted classes distribution - data are highly imbalanced.

4 Method proposal

5 Preparation

Before performing experiments, we preprocessed the data. At first, useless columns like id of transaction or those with too many missing values (more than 50%) are dropped. Our preprocessing pipeline can be devided into two branches - preprocessing of numeric and categorical attributes. In numerical attributes, missing values are filled in with mean value, then all values are normalized (some machine learning algorithms are sensitive to different scales). For categorical attributes,

[1] https://www.kaggle.com/c/ieee-fraud-detection/data

transformation of emails to email domains has been performed, missing values have been filled with most frequent values, small categories (smaller than 5%) merged into one *other* category and at the end, all features have been one-hot encoded. Also, missing values indicators have been added to both, numerical and categorical attributes.

6 Experiments

7 Conclusion

References

1. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Advances in Engineering Software **69**, 46–61 (Mar 2014). https://doi.org/10.1016/j.advengsoft.2013.12.007, https://www.sciencedirect.com/science/article/abs/pii/S0965997813001853