Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach

Roberto Saia and Salvatore Carta
Department of Mathematics and Computer Science
University of Cagliari, Italy

SECRYPT 2017, 14th International Conference on Security and Cryptography
24÷26 July 2017 - Madrid, Spain

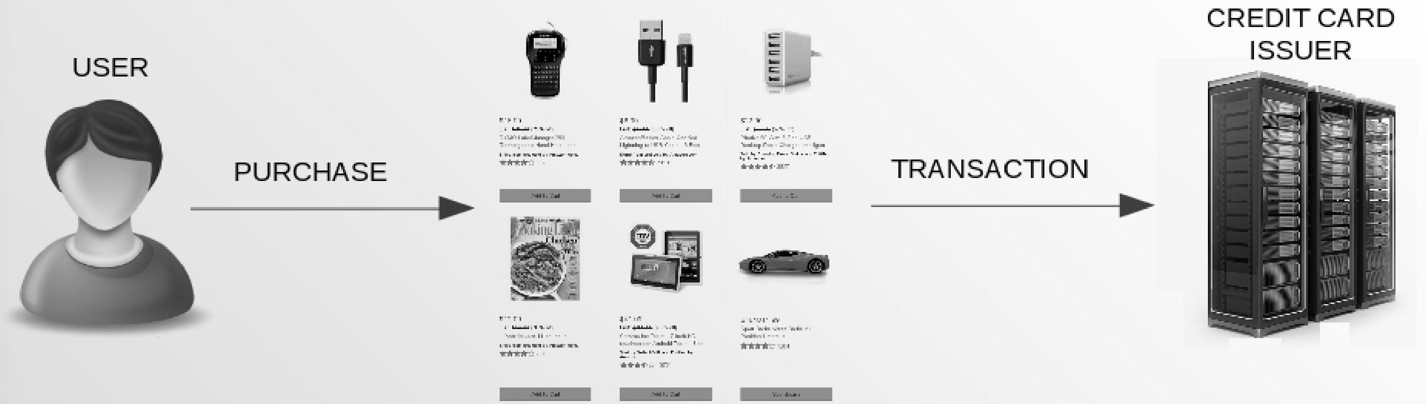
Outline

- **Overview**
  Introduction ▪ State of the Art ▪ Limits
- **Intuition**
  Idea ▪ Formalization ▪ Advantages
- **Implementation**
  Data Definition ▪ Data Evaluation ▪ Data Classification
- **Results**
  Experiments ▪ Conclusions ▪ Future Work

OVERVIEW

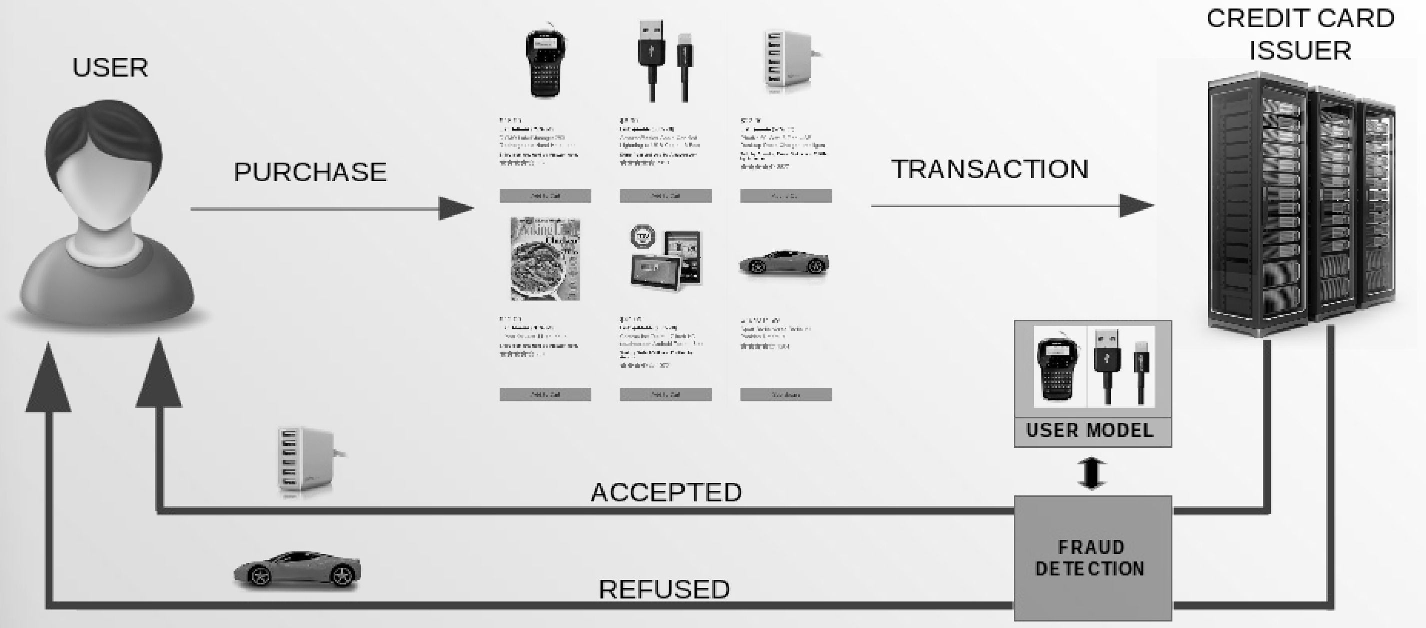
Introduction

- The main aim of a **Fraud Detection** task is the classification of a financial transaction as *legitimate* or *fraudulent*



Introduction

- It is usually performed on the basis of a **model** defined by using the previous transactions of a user



State of the Art

- Many state-of-the-art **approaches** have been designed to perform the ***Fraud Detection*** task, by exploiting several techniques, such as:
  - *Data Mining*[1]
  - *Artificial Intelligence*[2]
  - *Fuzzy Logic*[3]
  - *Machine Learning*[4]
  - *Genetic Programming*[5]
  - … and many more

State of the Art

- Regardless of the adopted technique, the definition of effective *Fraud Detection* models represents a **hard challenge** for several factors

- The most important of which is the **Imbalanced class distribution**[6]

Limits

- The **imbalanced class distribution** exists because the number of *fraudulent* cases is usually **much smaller** than that of the legitimate ones
- This reduces the **effectiveness** of the most widely used approaches[6]

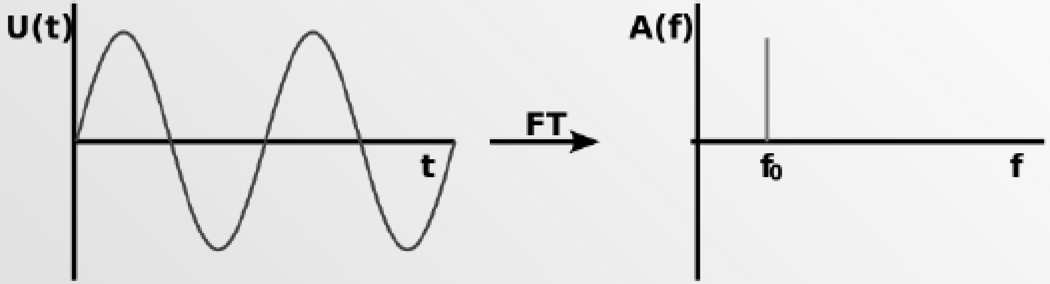
LEGITIMATE CASES **VS** FRAUDULENT CASES

INTUITION

Idea

- Overcome the **imbalanced class distribution** issue by using a single class of data (*legitimate transactions*) during the model definition

- Perform this operation through a model able to well characterize the legitimate transactions, defining it in the **frequency domain**



We move the data analysis from the canonical domain to the frequency one in order to obtain a **more stable** model

Formalization

- We propose a **frequency-domain-based pattern mining** process able to evaluate a new transaction in a novel way

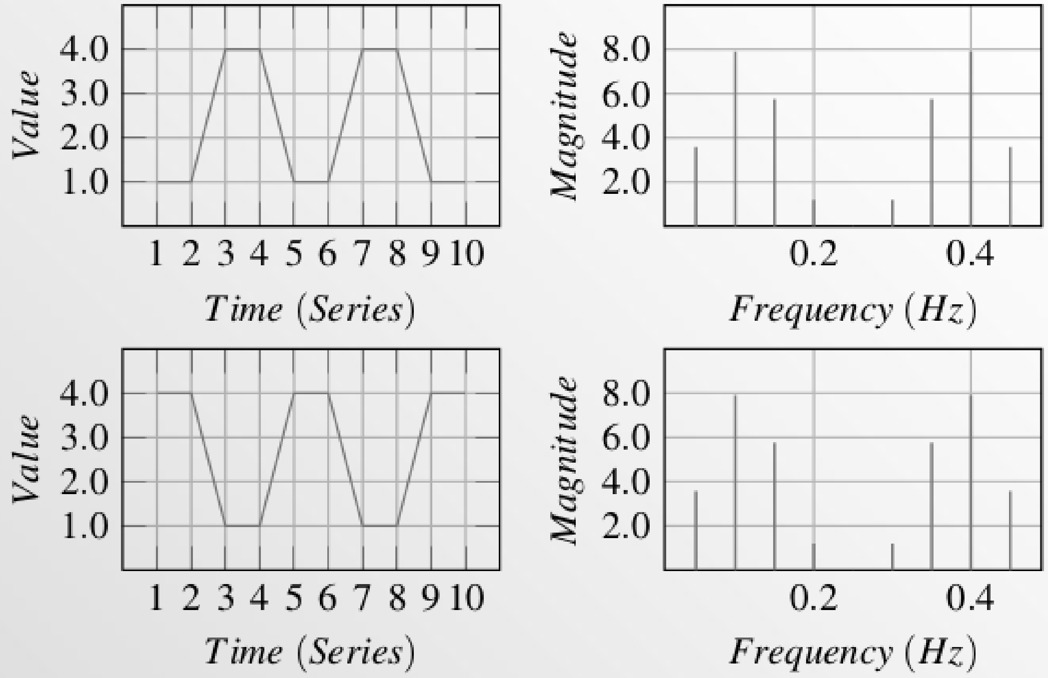
- It is made by exploiting the **Fourier transformation**[7], applying it on the values assumed by the **features*** of a transaction

- We perform this by considering these values as a **time series**

* It contains several information about the credid card transaction, e.g., **place**, **amount**, **date**, etc.

Formalization

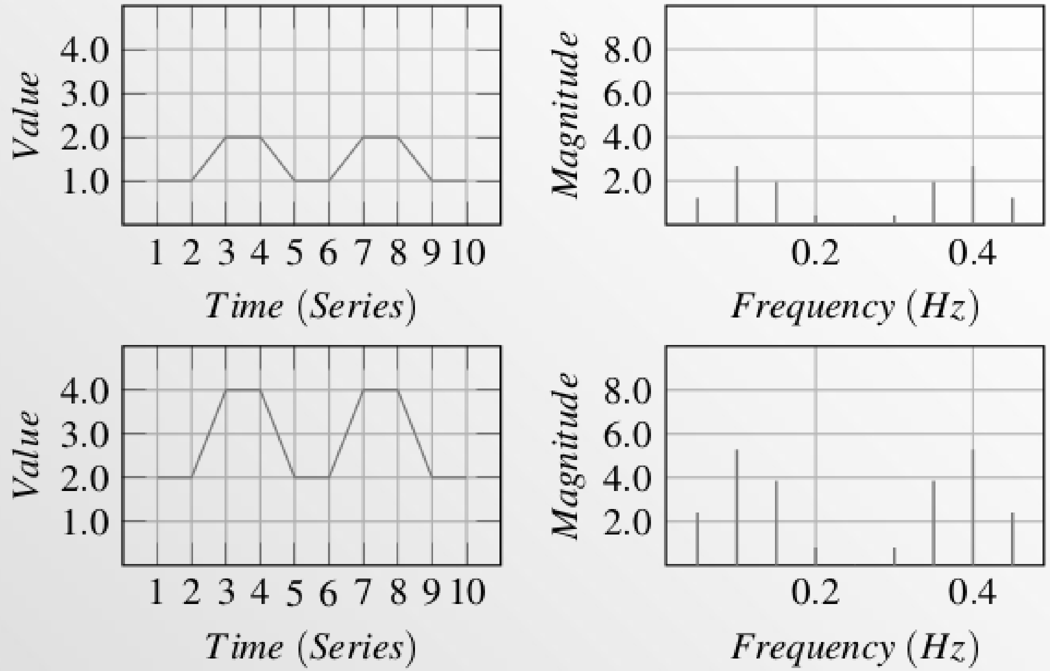
- The first interesting property of the new data representation is the **phase invariance**



- This allows us to detect **specific patterns**, also when they shift along the transaction features

Formalization

- The second interesting property of the new data representation is the **amplitude correlation**



- This grants that our model is able to **differentiate** same patterns with different feature values

Advantages

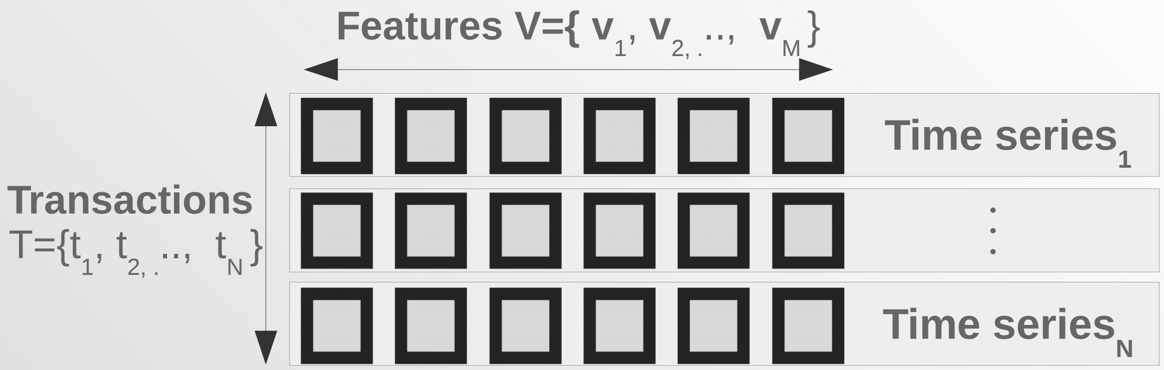
- The adoption of a model based on the frequency pattern leads toward some advantages:
  - It **overcomes** the imbalance class distribution issue by using only a class of data (previous legitimate transactions)
  - The use of only *legitimate* cases, allows us to operate in a **proactive** way
  - It also overcomes the problems related to the **cold-start** issue[8]

IMPLEMENTATION

Implementation

Step 1 of 3 - Data Definition

- Definition of the **time series** in terms of values assumed by the features of a transaction



- Each time series is processed by using a **Discrete Fourier Transform** process

Implementation

Step 2 of 3 – Data Evaluation

- Comparison of the **frequency patterns** of two transactions
- It is performed by measuring the difference **Δ** between the **magnitude** of each frequency component **f*** of the pattern

$$\Delta = \left(|f_x^1| - |f_x^2| \right), \ with \ |f_x^1| \geq |f_x^2|$$

*f_x^1 and f_x^2 denote, respectively, the same component of the two transactions

Implementation ■■■

Step 3 of 3 – Data Classification

- Definition of an **algorithm** able to classify a new transaction as *legitimate* or *fraudulent*, on the basis of the previous comparison process

- It counts how many times the **magnitude** of each frequency component of a new transaction is below or above that of each previous legitimate transaction

- The **classification** of the new transaction is made on the basis of the final result* of this operation

* A tolerance can be taken into account

RESULTS

Experiments

- We conducted the **experiments** by following the *k-fold cross validation* criterion, with *k=10*
- We used a **real-world dataset**(*) with 284,807 transactions and a *strong imbalanced* data distribution (only 492 fraud cases)
- The selected competitor was **Random Forests**, since it represents one of the most performing state-of-the-art approach in this field[11]

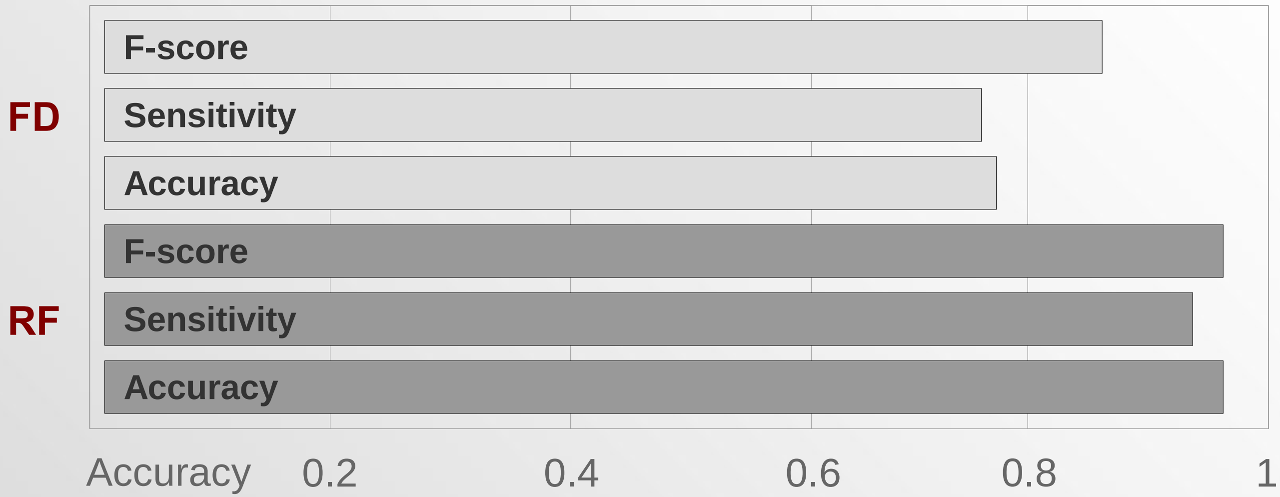
(*) https://kaggle.com/dalpozz/creditcardfraud

Experiments

- The *first three* experiments show the *general performance* achieved by our approach and Random Forests, in terms of ***Accuracy***, **Sensitivity**, and ***F-score***
- The *fourth* experiment is instead aimed to evaluate our predictive model in terms of **AUC**

Experiments

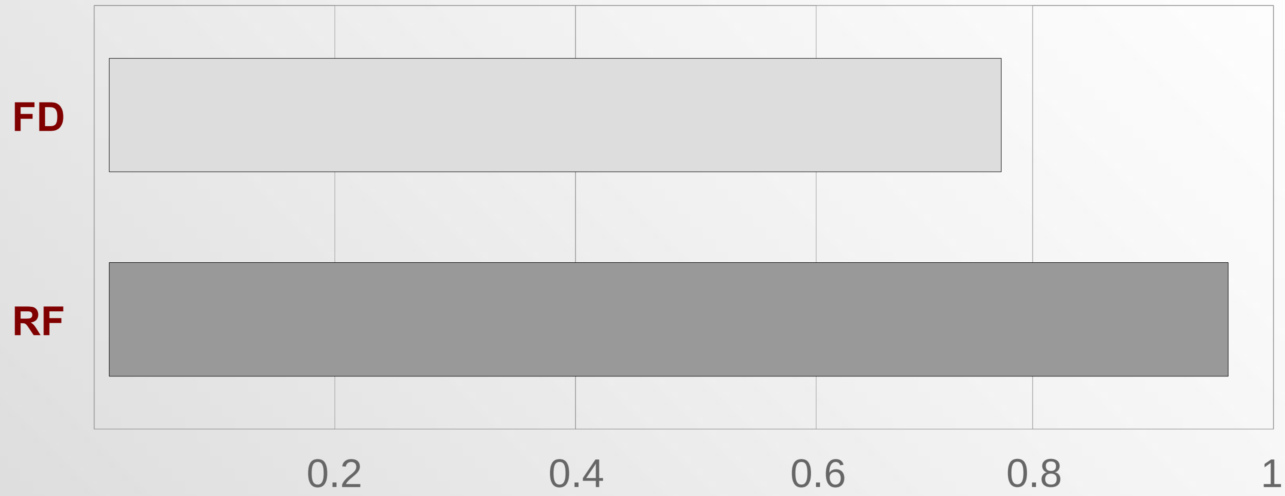
- The general performance of our **FD** approach in terms of ***Accuracy***, ***Sensitivity***, *and* ***F-score*** is close to that of *Random Forests* (**RF**)



- In spite of the fact that we did **<u>not use</u>** any past default instance to train our model

Experiments

- The last set of experiments measure the performance in terms of **AUC** (*Area Under ROC Curve*)



- Also in this case, the performance of our *predictive model* is close to that of **RF**

Conclusions

- **Fraud Detection** techniques cover a crucial role in many financial contexts
- The proposed approach allows a Fraud Detection system to face some well-known issues, such as the **unbalanced distribution of data** and the **cold-start**
- It is made by adopting a proactive strategy based on the **frequency pattern** of the transactions

Conclusions

- Such **proactive** approach does not replace the existing **retroactive** state-of-the-art approaches

- It is aimed to face some well-known problems, such as the **cold-start** and the **data unbalance** ones

- It can be used to define **hybrid** fraud detection approaches able to operate in all real-world scenarios

Future Work

- A future work will be the definition and evaluation of a **hybrid fraud detection approach** that exploits both previous *legitimate* and *fraudulent* cases

- We also consider to evaluate our approach in **heterogeneous scenarios** that involve different types of financial data, such as those that characterize the **E-commerce** environment

References

1) CLek, M., Anandarajah, B., Cerpa, N., and Jamieson R. (2001). *Data mining prototype for detecting e-commerce fraud.* Proceedings of the 9th European Conference on Information Systems, Global Cooperation in the New Millennium, ECIS 2001, Bled, Slovenia, June 27-29, 2001, pages 160–165.
2) Hoffman, A. J. and Tessendorf, R. E. (2005). *Artificial intelligence based fraud agent to identify supply chain irregularities*. In Hamza, M. H., editor, IASTED International Conference on Artificial Intelligence and Applications, part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria, February 14-16, 2005, pages 743–750. IASTED/ACTA Press.
3) Lenard, M. J. and Alam, P. (2005). *Application of fuzzy logic fraud detection*. In Khosrow-Pour, M., editor, Encyclopedia of Information Science and Technology (5 Volumes), pages 135–139. Idea Group.
4) Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C. C., and Albrecht, W. S. (2012). *Machine learning methods for detecting patterns of management fraud.* Computational Intelligence, 28(4):505–527.
5) Assis, C., Pereira, A. M., de Arruda Pereira, M., and Carrano, E. G. (2010). *Using genetic programming to detect fraud in electronic transactions*. In Prazeres, C. V. S., Sampaio, P. N. M., Santanch`e, A., Santos, C. A. S., and Goularte, R., editors, A Comprehensive Survey of Data Mining-based Fraud Detection Research, volume abs/1009.6119, pages 337–340.
6) Japkowicz, N. and Stephen, S. (2002). *The class imbalance problem: A systematic study*. Intell. Data Anal., 6(5):429–449.
7) Duhamel, P. and Vetterli, M. (1990). Fast fourier transforms: a tutorial review and a state of the art. Signal processing, 19(4):259–299.
8) Donmez, P., Carbonell, J. G., and Bennett, P. N. (2007). Dual strategy active learning. In ECML, volume 4701 of Lecture Notes in Computer Science, pages 116–127. Springer.

Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach

THANK YOU FOR YOUR ATTENTION