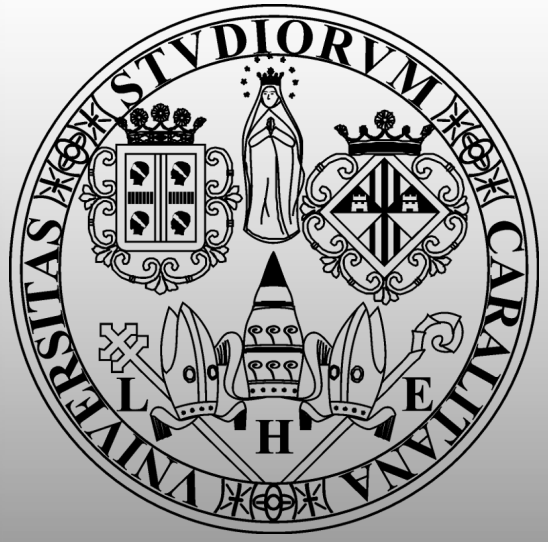


A Discrete Wavelet Transform Approach to Fraud Detection

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

NSS 2017, 11th International Conference on Network and System Security
International Workshop on Security Measurements of Cyber Networks (SMCN)
21÷23 August 2017 - Helsinki, Finland

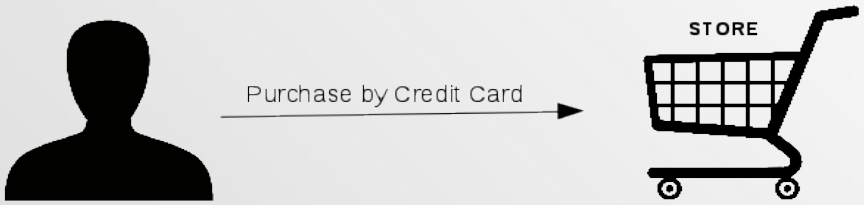
Outline

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

OVERVIEW

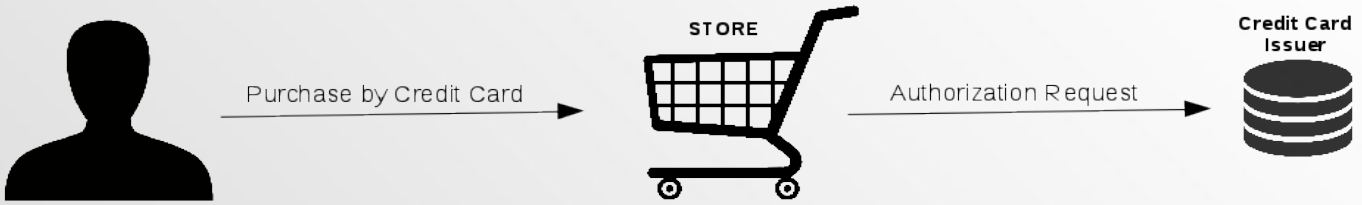
Introduction

- A **Fraud Detection** process is aimed to classify a new credit card transaction as *legitimate* or *fraudulent*



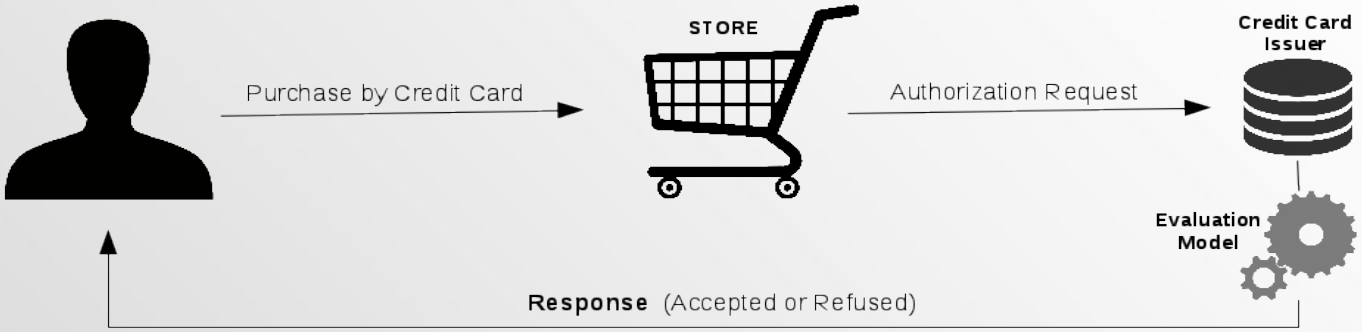
Introduction

- This happens when the **authorization request** arrives at the credit card issuer



Introduction

- The **response** depends on an **evaluation model** previously defined by using the user's past transactions



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
- Define this model in a **new domain** that allows us to well characterize this class of data
- This goal is achieved thanks to the **different point of view** on data offered by the new domain

Formalization

- We propose a **transformed-domain-based pattern mining** process able to evaluate a new transaction in a novel way
- It is made by exploiting the **Discrete Wavelet Transformation**[7], applying it on the values assumed by the transaction **features***
- We perform this operation by considering these values as a **time series**

* They contain several information about a credit card transaction, e.g., **place**, **amount**, **date**, etc.

Formalization

- The first interesting property of the new data representation is the **dimensionality reduction**
- The DWT process can reduce the time-series data, offering a **compact representation** that preserves the original information
- This allows a Fraud Detection system to reduce the **computational complexity** of the involved processes

Formalization

- The second interesting property of the new data representation is the **multiresolution analysis**
- The DWT process allows us to define **several time series** on the basis of the original one, preserving the information
- A Fraud Detection system can exploit this property to obtain two **different point of views** on data:
  - **approximated** (*data overview*)
  - **detailed** (*data changing*)

Formalization

- Our approach exploits **both** the aforementioned properties
- The original time series are transformed by using the **Haar wavelet***
- It offers us an approximated point of view on data that allows us to **model** the considered class of information

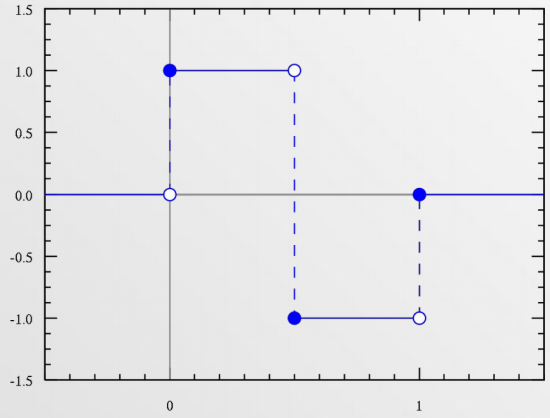
* The Haar wavelet represents the simplest possible wavelet.

Advantages

Haar wavelet is a sequence of rescaled *square-shaped* functions which represent a **wavelet family**

They are based on the following **Mother** and **Father** (*scaling*) functions:

$$\psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2}, \\ -1 & \frac{1}{2} \le t < 1, \\ 0 & \text{otherwise.} \end{cases} \qquad \varphi(t) = \begin{cases} 1 & 0 \le t < 1, \\ 0 & \text{otherwise.} \end{cases}$$



Advantages

Suppose we have a **time series** (**TS**) composed by **N** values (where **N** is even)

TS = (6, 12, 15, 15, 14, 12, 120, 116)

We calculate the **pair-wise average** $\mathbf{S}_k = (TS_{2k} + TS_{2k+1}) / 2$ for k=0, …, N/2 -1

TS = (6, 12, 15, 15, 14, 12, 120, 116) → **S** = (9, 15, 13, 118)

We also calculate the **directed distances** $\mathbf{D}_k = (TS_{2k} - TS_{2k+1}) / 2$ for k=0, …, N/2 -1

TS = (6, 12, 15, 15, 14, 12, 120, 116) → **D** = (-3, 0, 1, 2)

S and **D** allow us to recover the original information in **TS**

This process can be repeated **recursively**

and it allows us to define a **more stable** model of evaluation

Advantages

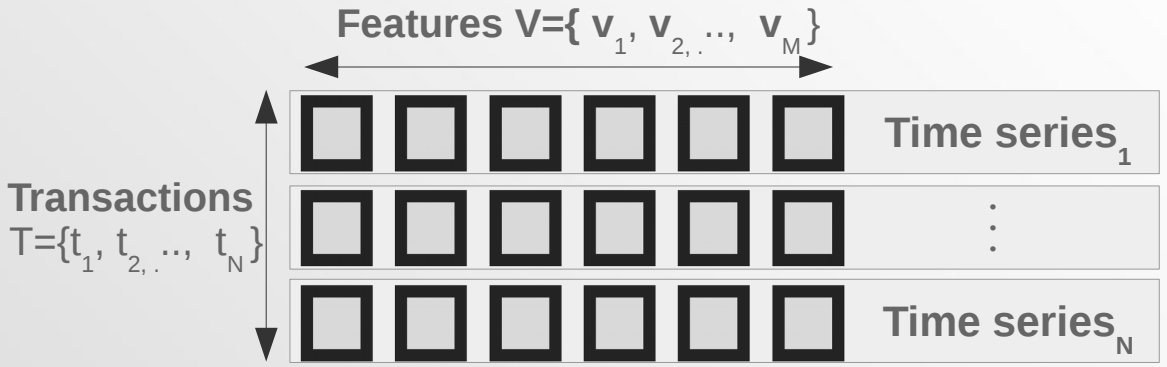
- The adoption of a model based on the wavelet leads toward some advantages:
  - It **overcomes** the imbalance class distribution issue by using only a class of data (previous legitimate transactions)
  - The use of only *legitimate* cases, allows us to operate in a **proactive** way
  - It also overcomes the problems related to the **cold-start** issue[9]

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 Wavelet Transform** process

Implementation

Step 2 of 3 – Data Evaluation

- Comparison of the **wavelet patterns** of two transactions
- It is performed in terms of **cosine similarity** between the output vectors of the wavelet processes

$$\Delta = Cosim(F(t), F(\hat{t})), \;\; with \;\; c = \begin{cases} \Delta \geq \alpha, & \text{legitimate} \\ \Delta < \alpha, & \text{fraudulent} \end{cases}$$

α is a threshold experimentally defined

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 takes into account the **cosine similarity** between the new transaction to evaluate and each previous legitimate transaction
- The **classification** of the new transaction is based on the average of these results

Require: T_+=Legitimate previous transactions, $\hat{t}$=Unevaluated transaction, α=Threshold
Ensure: β=Classification of the transaction $\hat{t}$

```
1: procedure TRANSACTIONEVALUATION(T+, t̂)
2:     ts1 ← getTimeseries(t̂)
3:     sp1 ← getDWT(ts1)
4:     for each t in T+ do
5:         ts2 ← getTimeseries(t)
6:         sp2 ← getDWT(ts2)
7:         cos ← cos + getCosineSimilarity(sp1, sp2)
8:     end for
9:     avg ← cos / |T+|
10:    if avg > α then β ← true else β ← false
11:    return β
12: end procedure
```

RESULTS

Experiments

- We conducted the **experiments** by adopting 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

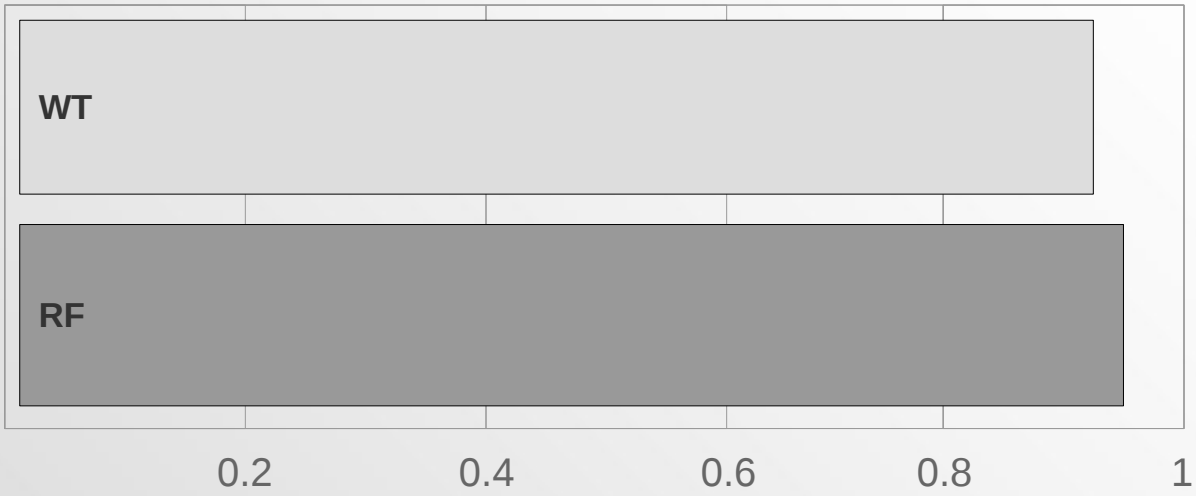
Experiments

- The *first series of* experiments shows the *general performance* achieved by our approach and Random Forests, in terms of ***F-score***
- The *second series of* experiment is instead aimed to evaluate our predictive model in terms of Area Under the ROC* Curve (**AUC**)

(*) Receiver Operating Characteristic

Experiments

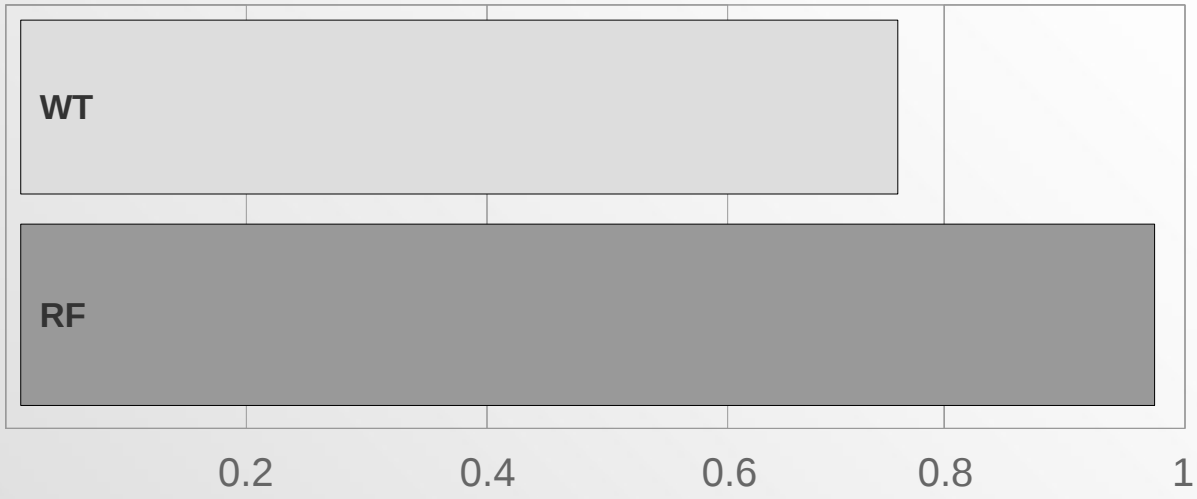
- The general performance of our **WT** approach in terms of ***F-score*** is close to that of *Random Forests* (**RF**)



- In spite of the fact that we did **not use** any past default instance to train the evaluation model

Experiments

- The last set of experiments measures the predictive model performance in terms of **AUC**



- Also in this case, the performance of our *predictive model* is quite 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 an evaluation model defined by exploiting the **Wavelet Transformation**

Conclusions

- Such **proactive** approach is not designed to replace the existing **retroactive** state-of-the-art approaches
- It is aimed to face some well-known problems that affect them, 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 are also considering the evaluation of our approach in **different scenarios** (e.g., those related to the *credit scoring* tasks)

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) Percival, D.B., Walden, A.T.: Wavelet methods for time series analysis. Volume 4. Cambridge university press (2006).

8) Lessmann, S., Baesens, B., Seow, H., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1):124–136.

A Discrete Wavelet Transform Approach to Fraud Detection

THANK YOU FOR YOUR ATTENTION

NSS 2017, 11th International Conference on Network and System Security
International Workshop on Security Measurements of Cyber Networks (SMCN)
21÷23 August 2017 - Helsinki, Finland