Unbalanced Data Classification in Fraud Detection by Introducing a Multidimensional Space Analysis

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

IoTBDS 2018, 3nd International Conference on Internet of Things, Big Data and Security
19÷21 March 2018 – Funchal, Madeira - Portugal

Outline

- **Overview**
  Introduction
  State of the Art
  Limits
- **Intuition**
  Idea
  Formalization
  Advantages
- **Implementation**
  Metrics
  Classification
- **Conclusions**
  Experimental Results
  Future Work

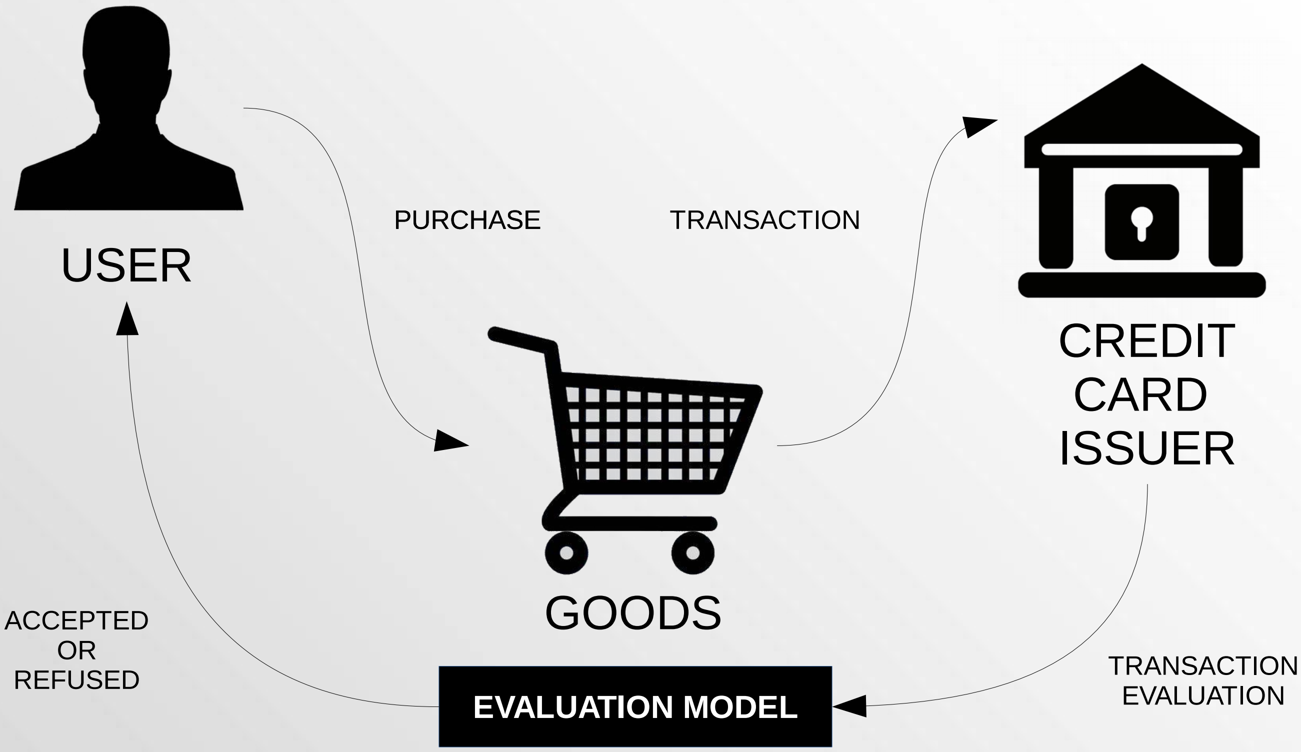
OVERVIEW

Introduction

A **Fraud Detection** task is aimed to classify a financial transaction as *legitimate* or *fraudulent*

Introduction

Such operation is performed by defining an **evaluation model** on the basis of the previous user transactions

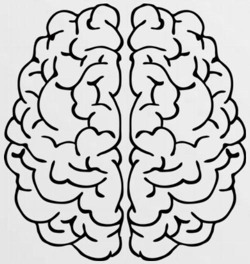


State of the Art

A large number of **state-of-the-art techniques** have been designed to address this problem

Machine Learning [3]

Artificial Intelligence [2]



Data Mining [1]

And so on . . .

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 **Data Imbalance** [4]

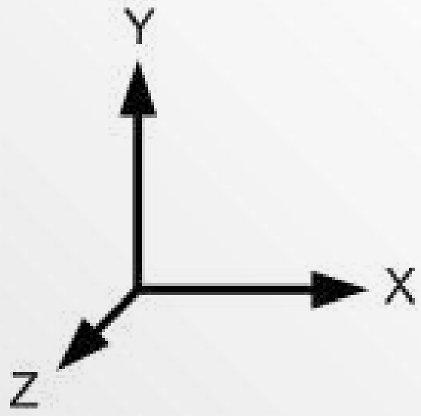
Limits

- **Data Imbalance** refers to a classification problem given by the fact that the classes are *not represented equally*
- Such configuration reduces the **effectiveness** of the state-of-the-art approaches [4]

INTUITION

Idea

- The main intuition on which the proposed approach relies is to perform the data analysis in a **three-dimensional space**, which is given by three different metrics of similarity



- It allows us to obtain a **better characterization** (wrt the state-of-the-art approaches) of each transaction in one of the two possible classes of destination (i.e., *legitimate* or *fraudulent*)

Formalization

- **Step 1 - Metrics Definition**: definition of three metrics aimed to compare two transactions in terms of different *similarity aspects*
- **Step 2 - Evaluation Criteria**: formalization of criteria used to evaluate a new transaction in the *three-dimensional space* given by the metrics
- **Step 3 - Algorithm Formulation**: formulation of an algorithm able to classify a new transaction as *legitimate* or *fraudulent*

Advantages

The adoption of a multidimensional similarity space approach leads toward some advantages:

- It offers a **better characterization** of each transaction in one of the two possible classes of destination (*legitimate* or *fraudulent*)
- It is able to operate in different real-world **data scenarios**, in terms of **size** and degree of **data imbalance**

IMPLEMENTATION

Metrics

The adopted similarity metrics are the following:

1. **Transactions Global Similarity Metric**: it measures the global similarity between two transaction vectors

2. **Features Local Similarity Metric**: it measures the similarity between transactions in terms of the weighted sequence of their features

3. **Features Global Similarity Metric**: it measures the global difference between two transactions in terms of their feature values

Metrics

They represent the **X**, **Y**, and **Z** dimensions:

X → 1

$$TGS(V_1, V_2) = \frac{V_1 \cdot V_2}{\|V_1\| \cdot \|V_2\|}$$

Y → 2

$$FLS(V^{(1)}, V^{(2)}) = TGS\left(idx\left(\overline{\overline{V^{(1)}}}\right), idx\left(\overline{\overline{V^{(2)}}}\right)\right)$$

with

$$\overline{\overline{V}} = \{|v_1| \leq |v_2| \leq \ldots \leq |v_M|\}$$

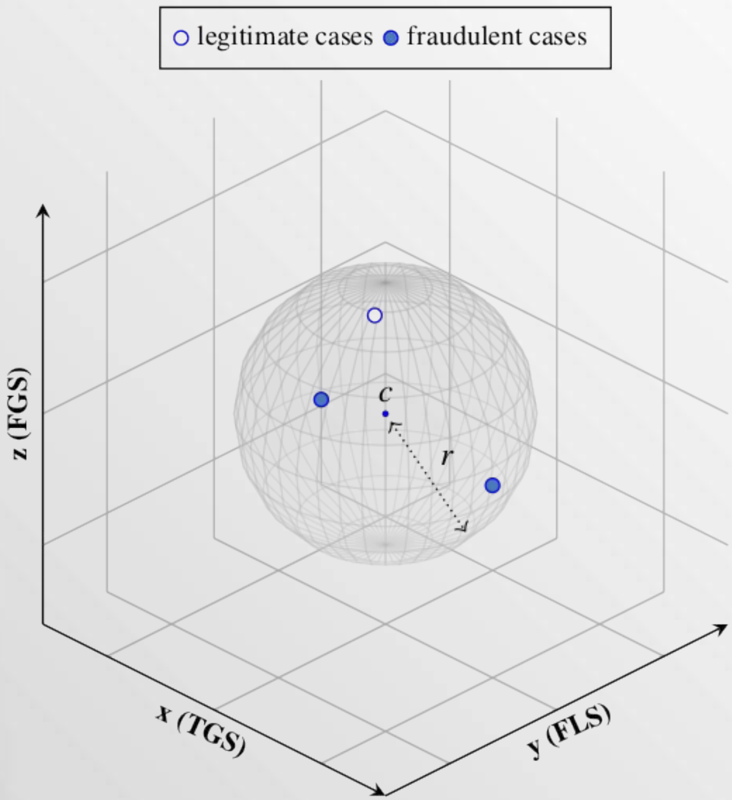
Z → 3

$$FGS(V^{(1)}, V^{(2)}) = 1 - \frac{\text{RMSE}}{max(\text{RMSE})}$$

with

$$\text{RMSE} = \sqrt{\sum_{m=1}^{M}\left(v_m^{(1)} - v_m^{(2)}\right)^2}$$

Classification



* The **radius** r is defined experimentally

1. We define a **center** c in our 3D space at the coordinates:

   ```
   max(TGS) − r
   max(FLS) − r
   max(FGS) − r
   ```

   by **comparing** the transaction to evaluate to the previous ones

2. The classification depends on the **nature** of the transactions bounded by the sphere of radius r* and center c

3. A transaction is **classified** as *legitimate* when the number of legitimate transactions bounded by the sphere is greater than that of the fraudulent ones, otherwise it is classified as *fraudulent*

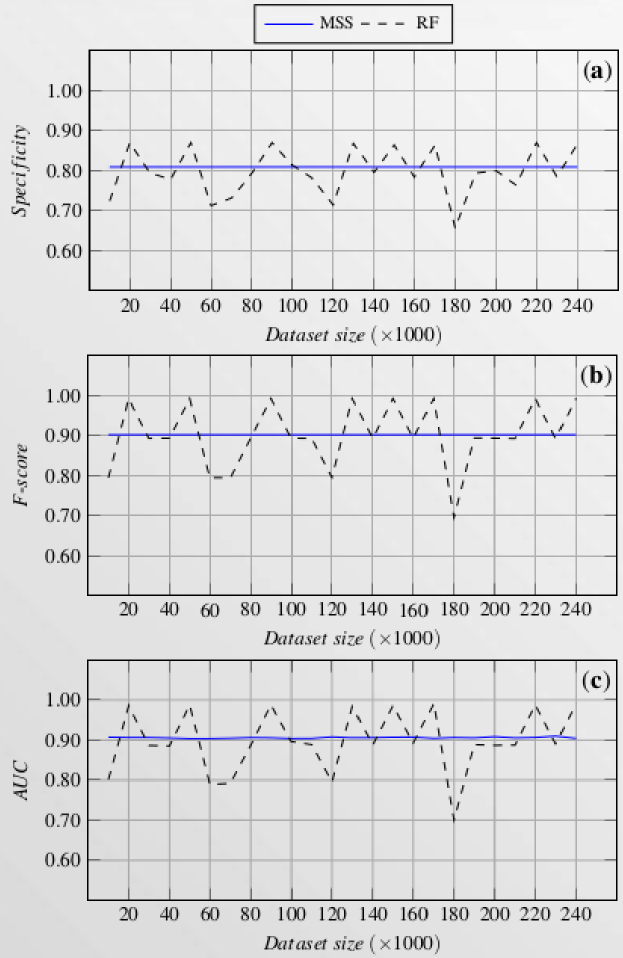
Classification

- The previous example shows a case where the evaluated transaction has been **classified as fraudulent**, because the number of fraudulent transactions bounded by the sphere of radius r and center c is greater than that of the legitimate ones

- In this process we adopt a **prudential criterion**, since the cases with equal number of legitimate and fraudulent transactions bounded by the sphere lead toward a fraudulent classification

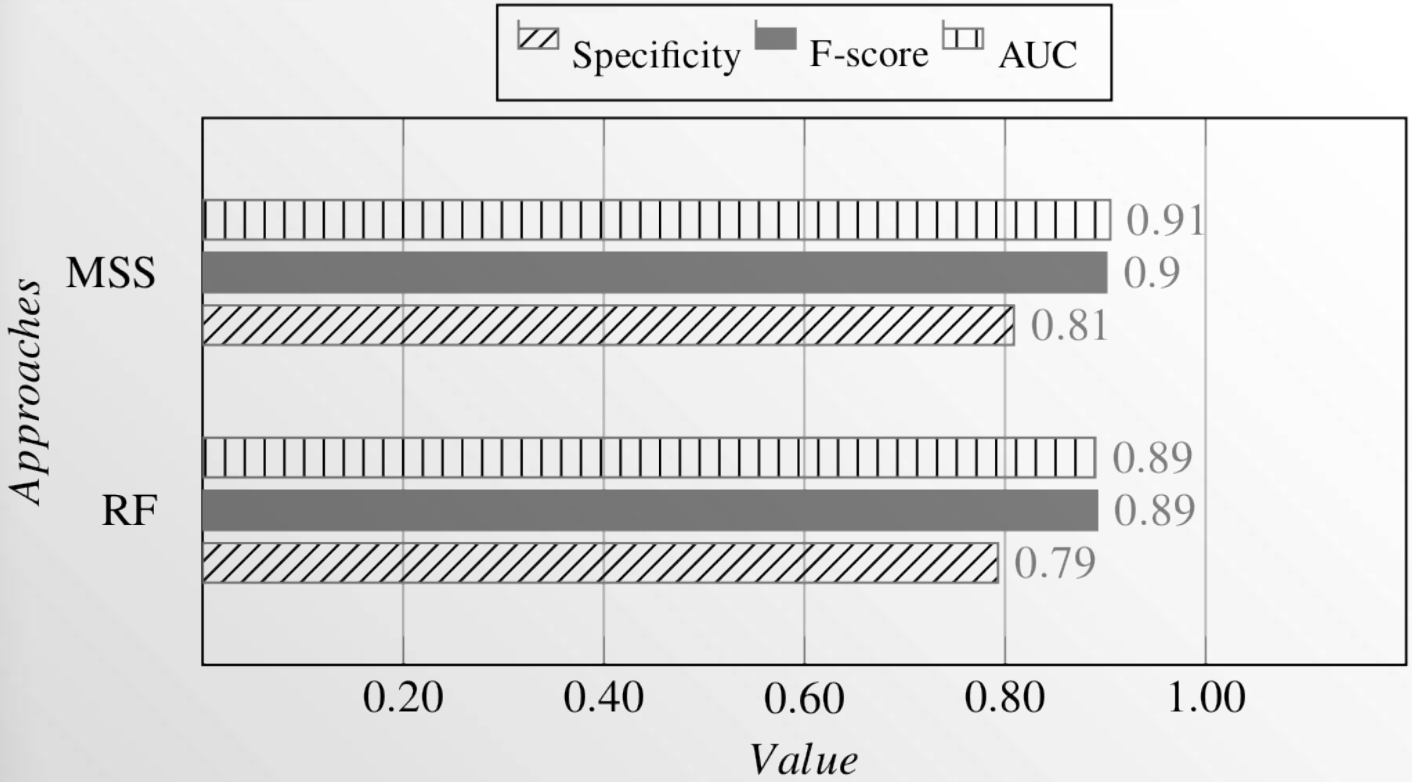
CONCLUSIONS

Experimental Results



- The first observation is related to the capability of our *Multidimensional Similarity Space* (**MSS**) approach to keep **constant** its performance, regardless of the size and the level of data imbalance, differently from its competitor Random Forests (**RF**)
- This mainly depends on the adopted strategy, which is able to better characterize the transactions in a 3D space of evaluation **less influenced** by the size and the level of data imbalance

Experimental Results



AVERAGE PERFORMANCE

Future Work

- Considering that the credit card fraud detection represents only one of the possible contexts where our approach can operate, a future work will be oriented to experiment it in other scenarios characterized by a high degree of data imbalance

- Another interesting future work would be the experimentation of additional metrics of similarity in order to improve the effectiveness of our classification approach

Unbalanced Data Classification in Fraud Detection by Introducing a Multidimensional Space Analysis

THANK YOU FOR YOUR ATTENTION

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) Japkowicz, N. and Stephen, S. (2002). *The class imbalance problem: A systematic study*. Intell. Data Anal., 6(5):429–449.