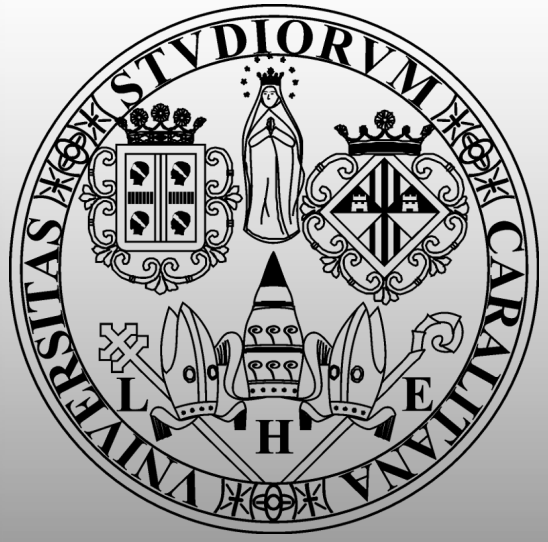


A Wavelet-based Data Analysis to Credit Scoring

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

ICDSP 2018, 2nd International Conference on Digital Signal Processing
International Academy of Computing Technology (IACT)
25÷27 February 2018 - Tokyo, Japan

Outline

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

OVERVIEW

Introduction

- The main aim of **credit scoring models** is the classification of the loan applications (from now on named as ***instances***) as *reliable* or *unreliable*

Introduction

- They are usually **designed** on the basis of the past loan applications
- They represent a **crucial** instrument for the financial operators

State of the Art

- Many **techniques** have been designed to address the *Credit Scoring* problem, by exploiting several models, such as:
  - *Statistical* and *Data Mining*
  - *Linear Discriminant*
  - *Logistic Regression*
  - *Neural Networks*
  - *Genetic Program*
  - *Decision-tree*
  - … and many more

State of the Art

- Regardless of the adopted technique, the definition of effective *Credit Scoring* models represents a **hard challenge** for several factors
- The most important of which is the **Imbalanced class distribution**[1]

Open Problems

- The **Imbalanced class distribution** exists because the number of un*reliable* instances (*non-refunded loans*) is **much smaller** than that of the *reliable* ones (*refunded loans*)
- This reduces the **effectiveness** of the most widely used approaches[2]

NON-DEFAULT CASES **VS** DEFAULT CASES

IDEA

Intuition

- Overcome the **imbalanced class distribution** issue by using a single class of data (reliable instances) 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 instance in a novel way
- It is made by exploiting the **Discrete Wavelet Transformation**[3], applying it on the values assumed by the **instances***
- We perform this operation by considering these values as a **time series**

* They contain information about the instance, e.g., applicant age, job, loan purpose, 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 Credit Scoring 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 Credit Scoring 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 class of information taken into account

* The Haar wavelet represents the simplest possible wavelet.

Advantages

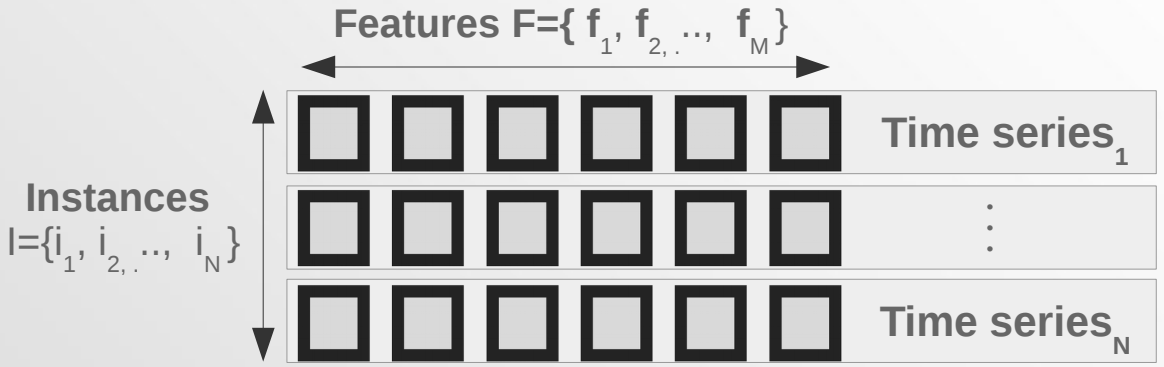
- The adoption of a model based on the wavelet leads toward some advantages:
  - It **overcomes** the imbalance class distribution issue by using only a class of data (previous reliable instances)
  - The use of only *reliable* cases, allows us to operate in a **proactive** way
  - It also reduces the problems related to the **cold-start** issue[4]

IMPLEMENTATION

Implementation

Input Data Definition

- Definition of the **time series** in terms of values assumed by the features of an instance



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

Implementation

Similarity Metrics Definition

We compare two instances in terms of *three metrics of similarity*:

- **Cosine Similarity** (Θ)
  Time-series vectors similarity

- **Features Weighted Sequence Similarity** (Φ)
  Time-series similarity in terms of weighted sequence of the features

- **Normalized Magnitude Similarity** (μ)
  Similarity in terms of magnitude between the DWT output of two time-series

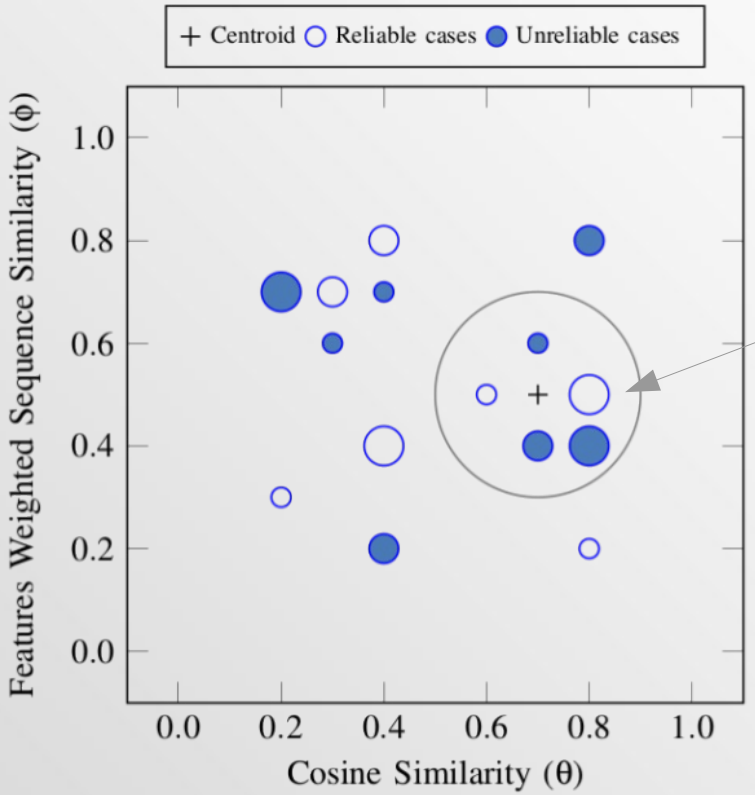
Implementation

Instance Classification

- We define a **radius (ρ)** value experimentally
- We define a **centroid** at coordinates **max(Cosine Similarity) − ρ** and **max(Features Weighted Sequence Similarity) − ρ**, (**max** *has been calculated between the instance to evaluate and all the instances in the training set*)
- A new instance is classified as reliable if the weight (in terms of **Normalized Magnitude Similarity**) of the selected *reliable* instances within the **radius ρ** is greater than that of the *unreliable* ones, otherwise it is classified as unreliable.

Implementation

Instance Classification



The size of circle represent s the *Normalized Magnitude Similarity* (**μ**)

In this case the instance under evaluation **is classified as unreliable**, since the sum of the weight of the **three unreliable instances** within the **radius** is **greater** than that of the other **two reliable ones.**

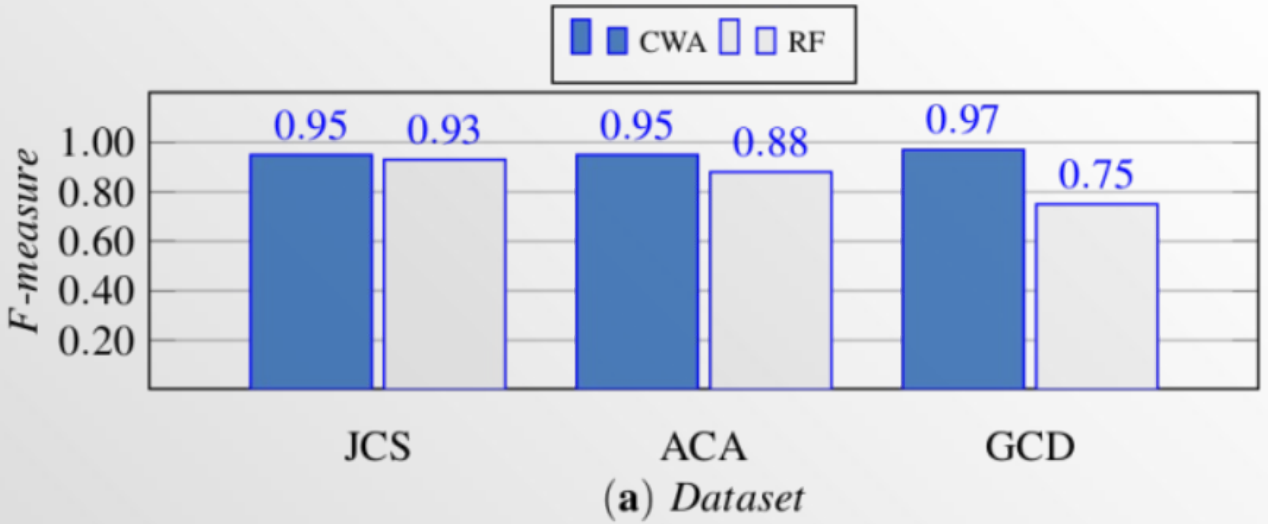
RESULTS

Experiments

- We conducted the **experiments** by using three **real-world datasets** that reproduce different data scenarios (for number of instances and class balancing)
- The selected competitor was **Random Forests**, since it represents one of the most performing state-of-the-art approach in this field[5]
- The first experiment evaluates the general performance (**F-measure**), while the second one evaluates our predictive model (**AUC**)

Experiments

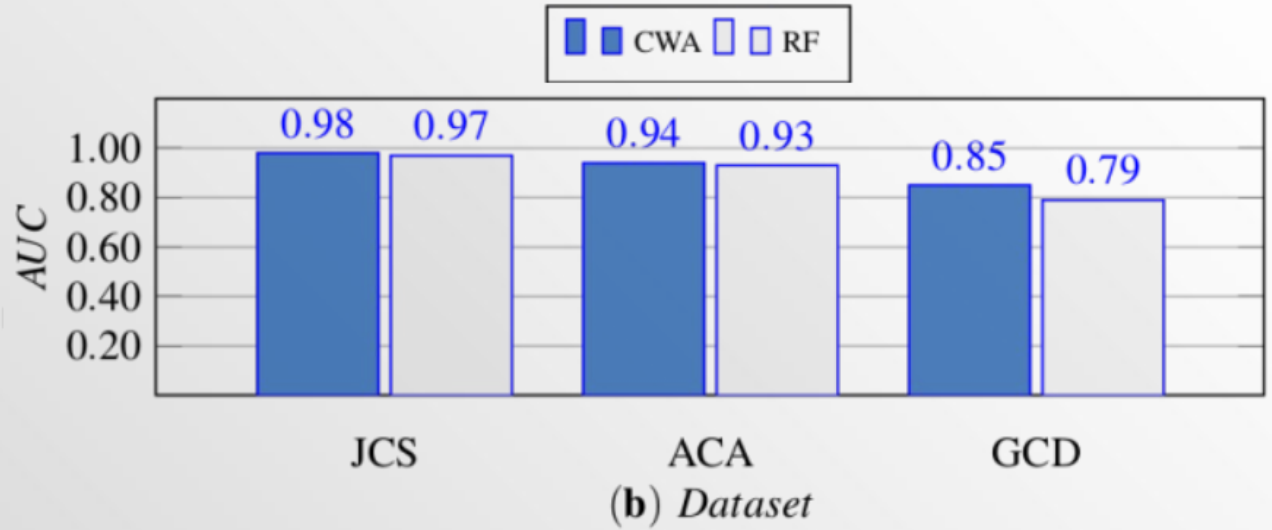
Our *Centroid Wavelet-based Approach* (**CWA**) ouperforms **RF** in terms of *F-measure*, regardless of data size and the level of imbalance



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

Experiments

- Also in terms of *AUC* our **CWA** approach outperforms its competitor **RF**



- It shows the **effectiveness** of its predictive model

Conclusions

- The **Credit Scoring** techniques cover a crucial role in many financial contexts
- The proposed approach allows a Credit Scoring 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 **Discrete 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 Credit Scoring approaches** able to operate in all real-world scenarios

Future Work

- Future work would be oriented to experiment **additional metrics** of similarity, as well as **other wavelet functions** in the *Discrete Wavelet Transformation* process, with the aim to further improve the effectiveness of the classification model

- Another interesting future work would be the experimentation of the proposed approach in **different scenarios**

References

[**1**] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. IEEE Trans. Knowl. Data Eng., 21(9):1263–1284.

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

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

[**4**] Attenberg, J. and Provost, F. J. (2010). Inactive learning?: difficulties employing active learning in practice. SIGKDD Explorations, 12(2):36–41.

[**5**] 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 Wavelet-based Data Analysis to Credit Scoring

THANK YOU FOR YOUR ATTENTION

ICDSP 2018, 2nd International Conference on Digital Signal Processing
International Academy of Computing Technology (IACT)
25÷27 February 2018 - Tokyo, Japan