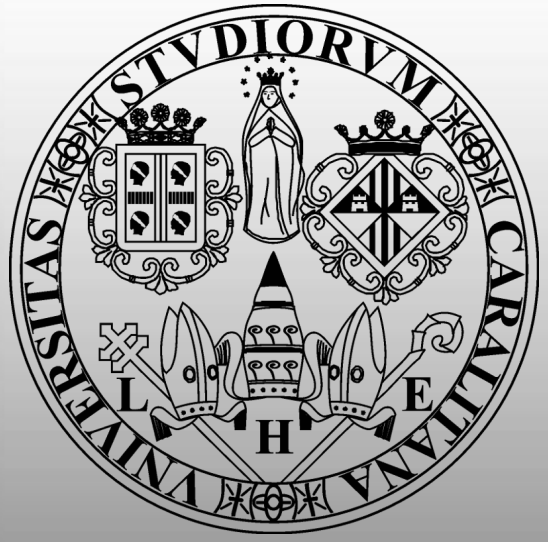


A Discretized Enriched Technique to Enhance Machine Learning Performance in Credit Scoring

Roberto Saia, Salvatore Carta, Diego Reforgiato Recupero, Gianni Fenu, and Marco Saia

Department of Mathematics and Computer Science
University of Cagliari, Italy

KDIR 2019, 11th International Conference on Knowledge Discovery and Information Retrieval
17÷19 September 2019 - Vienna, Austria

Outline

- **Overview**
  Introduction ▪ State of the Art ▪ Open Problems
- **Intuition**
  Idea ▪ Formalization
- **Implementation**
  Discretization ▪ Enrichment ▪ Instance Classification
- **Results**
  Experiments ▪ Conclusions and 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**

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

RELIABLE CASES **VS** UNRELIABLE CASES

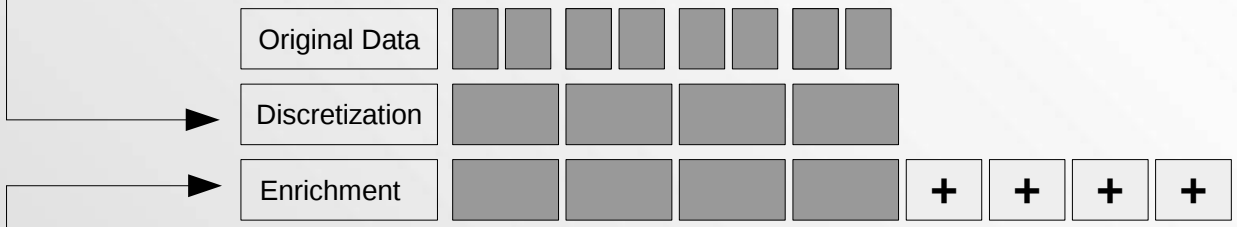
INTUITION

Idea

- The data unbalance issue can be reduced by improving the **instance** classes (i.e. reliable and unreliable cases) characterization
- We proposed a *Discretized Enriched Data* (**DED**) model able to face the heterogeneity of the event patterns
- This because the data domain is characterized by a **number of features** that could be very different, even when they define the same class of information

Formalization

Through a discretization process we **reduce** the instance patterns by grouping the similar ones in terms of feature values



Through an enrichment process we **improve** the instance characterization by adding a series of additional information

IMPLEMENTATION

Implementation

Step 1 of 3 - Discretization

On the basis of a η value experimentally defined, we perform the **data discretization** process $\xrightarrow{\Delta}$

$$\{f_1, f_2, \ldots, f_N\} \xrightarrow{\Delta} \{d_1, d_2, \ldots, d_N\}, \ \forall \ i \in I$$

Where f_1, f_2, ..., f_N denotes the original feature values of an instance i, and d_1, d_2, ..., d_N denotes these values after the discretization process

Implementation

Step 2 of 3 - Enrichment

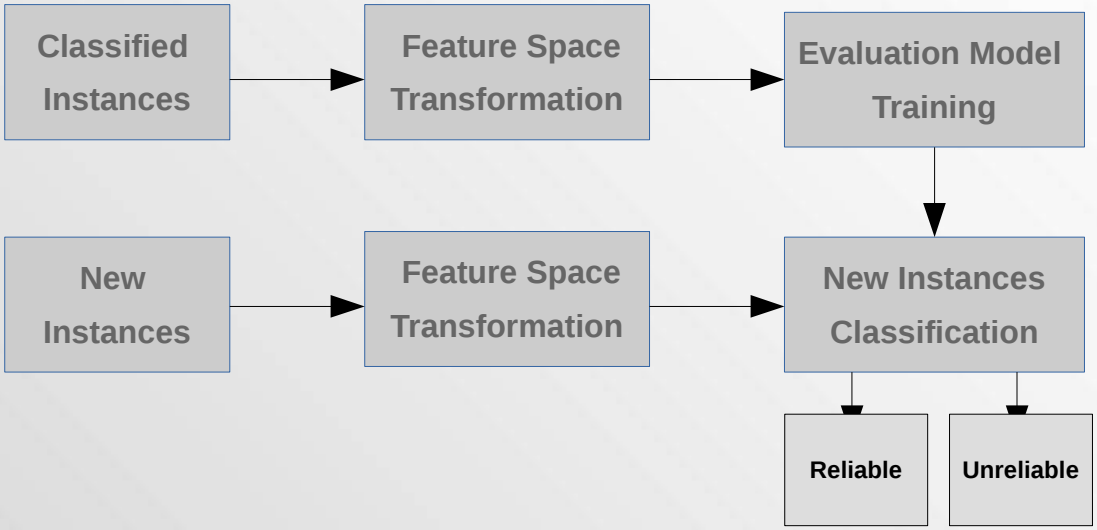
The discretized data are then enriched by adding to them the **Minimum** (m), **Maximum** (M), **Average** (A), and **Standard Deviation** (S) meta-information

$$\mu = \begin{cases} m = min(d_1, d_2, \ldots, d_N) \\ M = max(d_1, d_2, \ldots, d_N) \\ A = \frac{1}{N} \sum_{n=1}^{N} (d_n) \\ S = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (d_n - \bar{d})^2} \end{cases}$$

Implementation

Step 3 of 3 – Instance Classification

Our approach classifies the **new instances** as ***reliable*** or **unreliable** in the new feature space



RESULTS

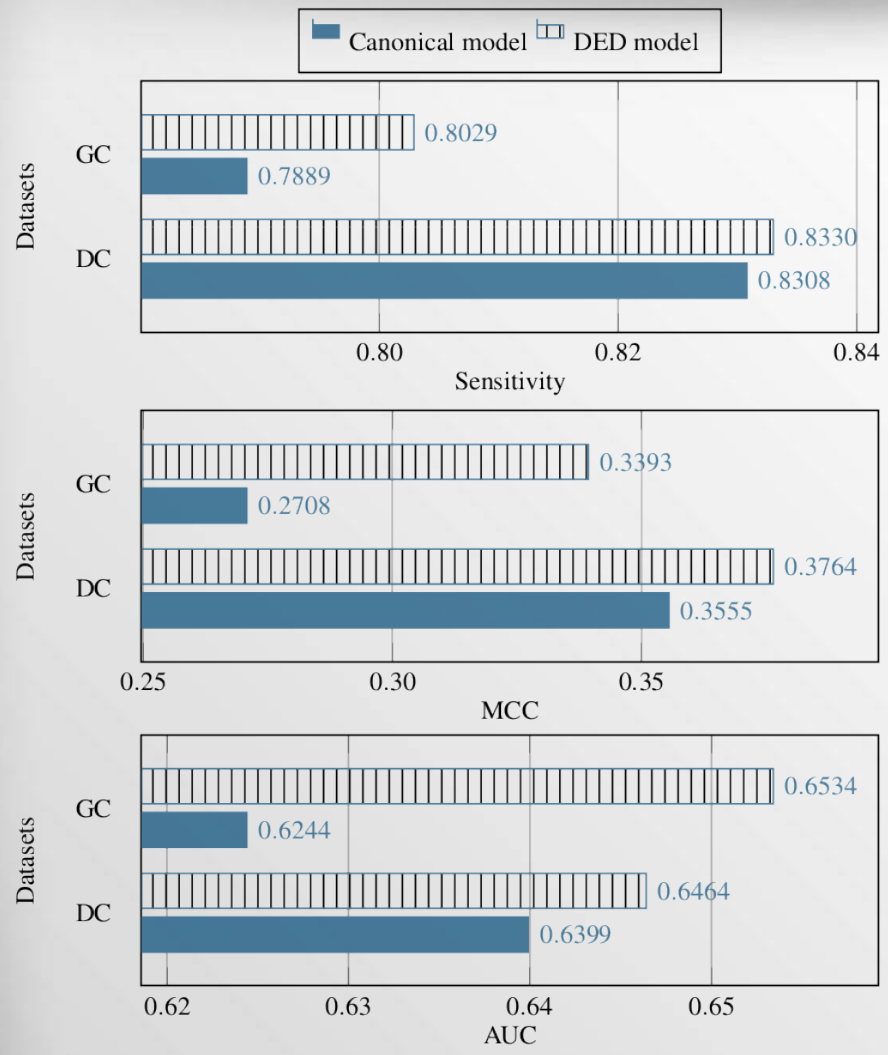
Experiments

- In addition to a canonical *k-fold cross validation* criterion, we **validated** our approach by dividing the dataset into two parts:
  - **In-sample part**, which has been used to select the internal algorithms parameters
  - **Out-of-sample part**, which has been used to evaluate the performance
- The *German Credit* (**GC**) and the *Default of Credit Card Clients* (**DC**) are the real-world dataset used in order to evaluate the performance of the proposed approach
- Our competitor, experimentally selected on the basis of its performance, is **Gradient Boosting**

Experiments

- We used the *50%* of each dataset as **in-sample** data, using the remaining *50%* as **out-of-sample** data
- We evaluated our predictive model in terms of **Sensitivity**, **Matthews Correlation Coefficient**, and **AUC**

Experiments



The results show that our **DED** model **outperforms** the canonical one in terms of *Sensitivity*, *MCC*, and *AUC* metrics, in both datasets

Conclusions

- The proposed model proved that the transformation of the original feature space, made by applying a discretization and an enrichment process, improves the performance of one of the most performing machine learning algorithm (i.e., Gradient Boosting)
- A possible **future work** would be the experimentation of such a model in the context of different classification algorithms configured through an ensemble strategy

A Discretized Enriched Technique to Enhance Machine Learning Performance in Credit Scoring

THANK YOU FOR YOUR ATTENTION

KDIR 2019, 11th International Conference on Knowledge Discovery and Information Retrieval
17÷19 September 2019 - Vienna, Austria