A Wavelet-based Data Analysis to Credit Scoring

Roberto Saia, Salvatore Carta, Gianni Fenu
Department of Mathematics and Computer Science
University of Cagliari, Via Ospedale 72 - 09124 Cagliari, Italy
Email: {roberto.saia, salvatore, fenu}@unica.it

ABSTRACT
Nowadays, the dramatic growth in consumer credit has made ineffective the methods based on the human intervention, aimed to assess the potential solvency of loan applicants. For this reason, the development of approaches able to automate this operation represents today an active and important research area named Credit Scoring. In such scenario it should be noted how the design of effective approaches represents an hard challenge, due to a series of well-known problems, such as, for instance, the data imbalance, the data heterogeneity, and the cold start. The Centroid wavelet-based approach proposed in this paper faces these issues by moving the data analysis from its canonical domain to a new time-frequency one, where this operation is performed through three different metrics of similarity. Its main objective is to achieve a better characterization of the loan applicants on the basis of the information previously gathered by the Credit Scoring system. The performed experiments demonstrate how such approach outperforms the state-of-the-art solutions.

CCS Concepts
•Information systems → Data stream mining; Clustering and classification; Business intelligence •Theory of computation→Pattern matching •General and reference → Metrics;

Keywords
Business intelligence; credit scoring; pattern mining; data processing; wavelets; classifications; metrics.

1. INTRODUCTION
The approaches of *Credit Scoring* are aimed to evaluate the user reliability in several contexts such as, for instance, those related to the loan applications (from now on named as *instances*). They cover a more and more crucial role in this our age dominated by the consumer credit, since the amount of money lost by the financial operators due to loans fully or partially not repaid depends on their effectiveness.

A *Credit Scoring* approach works by classifying each new instance as *reliable* or *unreliable* by exploiting an evaluation model defined on the basis of the previous instances. They can be considered as a series of statistical methods aimed to evaluate the probability that a new instance will lead to a fully or partially non-repayment of a loan [1]. The definition of effective *Credit* Scoring approaches is not an easy task, due to a series of well-known problems.

The idea around which this paper revolves is mainly based on the the exploitation of the *Discrete Wavelet Transformation* (DWT) [5], which is used in order to move the data analysis in a non-canonical time-frequency domain. In such domain the analysis is performed through three different metrics of similarity, which are aimed to better characterize the classes of data involved in the *Credit Scoring* processes. The contributions given by this paper are as follows:

- definition of the *time series* to be use as input of the DWT process, defined on the basis of the previous instances;
- conversion of the instance *time series* into the frequency-time domain by using the DWT process;
- formalization of the *Centroid Wavelet-based Approach* (CWA) able to classify new instances.

The paper is organized as follows: *Section II* provides an overview of the *Credit Scoring* scenario; *Section III* introduces the formal notation adopted in this paper; *Section IV* describes our approach; *Section V* provides details about the performed experiments and their results; *Section VI* draws certain conclusions and points to some further directions for research.

2. BACKGROUND AND RELATED WORK
An ideal *Credit Scoring* system should be able to evaluate each new instance correctly, by classifying it as *reliable* or *unreliable* on the basis of the information available in the previous instances. Literature proposes a considerable number of classification techniques aimed to perform such task [10], as well as many studies focused on the evaluation of their performance [4], on the optimal tuning of their parameters [2], and on the most suitable metrics of evaluation [7]. On the basis of the results offered by the *Credit Scoring* techniques is possible to predict when an application (e.g., for a loan) potentially leads towards a risk of partial or total non-repayment [13].

Regardless of the adopted technique, there are a number of problems that complicate such tasks. The imbalanced class distribution of data is the most important of them and it happens because the previous instances, collected by a *Credit Scoring* system to train its evaluation model, are composed by a big number of *reliable* cases, compared to the number of *unreliable* ones. It leads towards a reduction of the *Credit Scoring* techniques effectiveness [8]. Another problem to face is the *data heterogeneity*, which in literature is described as the incompatibility among similar features resulting in the same data being represented differently in different datasets. The *cold start* problem instead happens when the previous instances are not

ICDSP 2018, February 25–27, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6402-7/18/02…$15.00
DOI: https://doi.org/10.1145/3193025.3193026

representative for both classes of information (*reliable* and *unreliable*), preventing the definition of an effective evaluation model.

The proposed approach is mainly based on the *Discrete Wavelet Transformation* (DWT) process [12]. Such process exploits the *wavelets*, a task that in literature is usually performed in order to reduce the size or the noise of data (e.g., in the image compression and filtering tasks). The *wavelets* are mathematical functions that work by decomposing the input data into different frequencies at different scales. The input data of a DWT process is usually a *time series*, a sequence of values obtained by measuring the variations during the time of a specific type of data (e.g., voltage, temperature, etc.). The output is a new representation of data in a frequency-time domain (data representation in terms of both frequency and time). In our case, the *time series* used as input of the DWT process are the values assumed by the instance features.

The frequency-time domain offers us some interesting advantages, the most important of them is the *Multi-Resolution Analysis* operates by DWT that allows us to observe the data at different levels of resolutions [11], with the possibility of obtaining an approximated or detailed vision on them. Our approach exploits this in order to have a better characterization of the *reliable* and *unreliable* instances.

The *Equation 1* shows the formalization of the *Continuous Wavelet Transform* (CWT), where $\Psi(t)$ is the *mother wavelet* (i.e., a continuous function in both the time and frequency domain) and * denotes the complex conjugate.

$$X_W(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{+\infty} x(t) \Psi^* \left(\frac{t-b}{a} \right) dt \quad (1)$$

$$x_{m,n}(t) = \frac{1}{\sqrt{a^m}} \Psi \left(\frac{t - nb}{a^m} \right) \quad (2)$$

$$\Psi(t) = \begin{cases} 1, 0 \le t > 1/2 \\ -1, 1/2 \le t > 1 \\ 0, otherwise \end{cases} \quad \varphi(t) = \begin{cases} 1, 0 \le t > 1 \\ 0, otherwise \end{cases} \quad (3)$$

Considering that, for several reasons (e.g., the computational load), it is not possible to perform a data analysis by using all the wavelet coefficients, a common approach is to use a discrete subset of the upper half-plane, so we can be able to rebuild the original data by using the corresponding wavelet coefficients. Such discrete subset is composed by all the points (a^m, na^mb), where $m, n \in Z$, and after this operation we can formalize the *child wavelets (Equation 2)* .

A data compression that allows us to have a data overview (approximated data view) is related to the use of small scales, as this is equivalent to using high frequencies (because the scale is given by the formula *1/frequency*). In a opposite way, a data expansion that allows us to observe the data changing (detailed data view) is related to the use of large scales, as it is equivalent to using low frequencies.

A number of functions can be used as *mother wavelet* (e.g., *Haar*, *Daubechies*, *Symlets*, *Meyer*, *Coiflets*, etc), but for the objectives of this paper we take into account only the *Haar* [11] one. It is a sequence of rescaled square-shaped functions which represent a wavelet family. They are based on the *mother* Ψ and *father* φ (scaling) functions shown in *Equation 3*.

The proposed *Centroid Wavelet-based Approach* compares the instances in the new time-frequency domain through three metrics of similarity, which are aimed to evaluate different aspects of the instances.

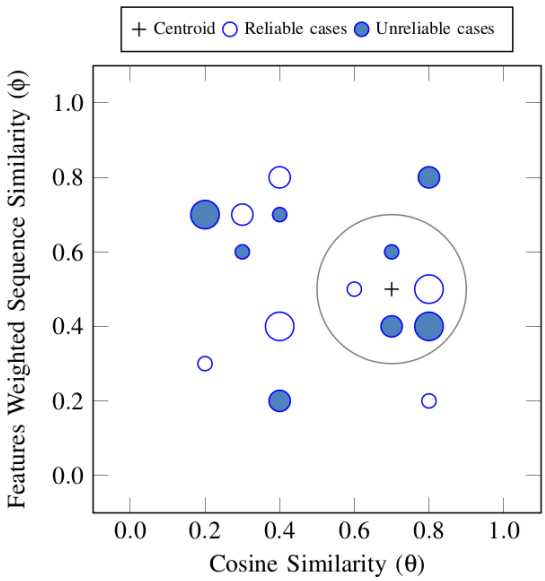

Figure 1. Evaluation Criterion

3. FORMAL NOTATION

Given a set of $I = \{i_1, i_2, \ldots, i_N\}$ classified instances , and a set of features $F = \{f_1, f_2, \ldots, f_M\}$ that compose each instance *i*, we denote as $I_+ \subseteq I$ the subset of *reliable* instances, as $I_- \subseteq I$ the subset of *unreliable* ones, and as *C={reliable,unreliable}* the set of possible instance classifications. It should be noted that an instance can belong only to one class $c \in C$. We also denote as $\hat{I} = \{\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_U\}$ a set of unclassified instances and as $E = \{e_1, e_2, \ldots, e_U\}$ these instances after the classification process, thus $|\hat{I}| = |E|$. Finally, we denote as $TS = \{ts_1, ts_2, \ldots, ts_Y\}$ and $O = \{o_1, o_2, \ldots, o_X\}$, respectively, input and output of the DWT process.

4. PROPOSED APPROACH

The proposed approach has been implemented through the three steps listed and explained in the following:

- **Input Data Definition**: definition of the *time series* to use in the DWT process, made by using the sequence of values assumed by each single feature of an instance;
- **Output Data Generation**: generation of the new data representation in the frequency-time domain, by processing the *time series* through the DWT process;

- **Instance Classification**: formalization of the *Centroid Wavelet-based Approach* (CWA) able to classify a new instance as *reliable* or *unreliable* on the basis of three different metrics of similarity.

4.1 Input Data Definition

The first step is aimed to prepare the *time series* to use as input in the DWT process. It is performed by using the sequence of values assumed by the features (set *F*) of an instance.

4.2 Output Data Generation

The second step performs the DWT process by using as input the *time series* in the $TS_{(I)}$ and $TS_{(I)}$ sets. Our approach exploits two wavelet properties. The first one is the *Dimensionality reduction*: the DWT process reduces the dimensionality of a *time series* by performing an orthonormal transformation that allows us to recover the original data. This can be exploited in order to reduce the computational load. The second one is the *Multiresolution analysis*: the DWT process allows us to analyze the data by using an approximated or detailed point of view.

4.3 Instance Classification

Each unevaluated instance $i \in I$ is classified after a comparison process performed between it and all the instances in the training set (i.e., the *reliable* ones in the subset I_+ and the *unreliable* ones in the subset I_-). Such comparison process is performed in terms of the three metrics detailed described later, i.e., the *Cosine Similarity* (Θ), the *Features Weighted Sequence Similarity* (Φ), and the *Normalized Magnitude Similarity* (μ). Premising that the *radius* (ρ) is a value experimentally defined in *Section V*, we classify a new instance by adopting the following criteria:

- first we define a *centroid* + (see *Figure 1*), using as coordinates of the *x* and *y* axes, respectively, *max(Θ)-ρ* and *max(Φ)*-ρ, where *max(Θ)* stands for the maximum value of the *Cosine Similarity* and *max(Φ)* stands for the maximum value of the *Features Weighted Sequence Similarity,* both calculated between the instance to evaluate and all the instances in the training set;
- the classification process is based on the type and magnitude of the instances bounded by the circular area of *radius* ρ, centered in + (the circular area of *Figure 1*, where the size of the cases represents the *Normalized Magnitude Similarity* μ of the instance);
- a new instance is classified as *reliable* if the weight (in terms of μ) of the selected *reliable* instances within the *radius* ρ is greater than that of the *unreliable* ones, otherwise it is classified as *unreliable*.

Figure 1 shows a case when 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. By following a prudential criterion, we classify the new instances as *unreliable* when the sum of the weight of the *reliable* ones within the *radius* is equal than that of the *unreliable* ones.

4.4 Metrics

This section describes the three metrics used in our approach.

Cosine Similarity: The *Cosine Similarity* (Θ) metric is able to measure the similarity between two vectors v_1 and v_2 with size larger than zero. More formally, given two vectors v_1 and v_2 of attributes, it is represented using a dot product and magnitude as shown in the *Equation 4*. We normalized the result in a range *[0,1]*, where *0* indicates two completely different vectors and *1* two equal vectors, and the intermediate values indicate different levels of similarity between the two vectors.

$$\theta(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|} \tag{4}$$

Features Weighted Sequence Similarity: The *Features Weighted Sequence Similarity* (Φ) is not a canonical metric, since it was defined in the context of this paper in order to evaluate the similarity between instances in terms of the weighted sequence of the features that composed them. The idea behind this metric is that similar instances present similar sequences of features, i.e., if we sort the features of two instances on the basis of their values, the sequences of their indexes will be similar in terms of *cosine similarity* Θ. More formally, given two instances $i^{(1)}$ and $i^{(2)}$ we calculate Φ as shown in *Equation 5*, where $TS^{(1)}$ and $TS^{(2)}$ are the *time series* of the instances to compare and the function *idx* gives us the sorted *time series TS* in terms of former element indexes (i.e., the indexes of the *TS* elements before sorting).

$$\Phi(TS^{(1)}, TS^{(2)}) = \theta\left(idx\left(\overline{\overline{TS^{(1)}}}\right), idx\left(\overline{\overline{TS^{(2)}}}\right)\right) \tag{5}$$

$$with \qquad \overline{\overline{TS}} = \{|ts_1| \leq |ts_2| \leq \ldots \leq |ts_Y|\}$$

Normalized Magnitude Similarity: The *Normalized Magnitude Similarity* (μ) is also a non-canonical metric defined in the context of this paper with the aim to evaluate the difference in terms of magnitude between two instances $i^{(1)}$ and $i^{(2)}$.

It is measured by taking into account their DWT outputs $O^{(1)}$ and $O^{(2)}$. It is calculated as shown in *Equation 6*, where *max(Δ)* is the maximum value assumed by Δ in the context of all the instances in the training set *I*.

$$\mu(O^{(1)}, O^{(2)}) = 1 - \frac{\Delta}{max(\Delta)}$$

$$with \tag{6}$$

$$\Delta = \sqrt{\sum_{0=1}^{X} (o^{(1)} - o^{(2)})^2}$$

5. EXPERIMENTS

Our approach has been developed in Java, by using the *Jwave*[1] library to perform the *Discrete Wavelet Transformations* and the WEKA[2] library to implement the state-of-the-art competitor used to evaluate its performance (i.e., *Random Forests*).

5.1 Datasets

The real-world datasets used to evaluate our approach are freely downloadable at the *UCI Repository of Machine Learning*

[1] https://github.com/cscheiblich/JWave/
[2] http://www.cs.waikato.ac.nz/ml/weka/

Databases[3]. They represent three benchmarks in the *Credit Scoring* field, allowing us to evaluate our approach in different data scenarios, both for number of instances and for class balancing.

German Credit Data (GCD): This dataset contains the classification of people as *reliable* or *unreliable* in terms of credit risks. We used the numerical version, which is composed by *1,000* instances: *700* classified as *reliable* (*70.0%*) and *300* classified as *unreliable* (*30.0%*). Each instance is defined by *24* features and a binary class variable (*reliable* or *unreliable*).

Australian Credit Approval (ACA): This dataset contains credit card applications classified as *reliable* or *unreliable* on the basis of their final outcome. We used the numerical version, which is composed by *690* instances: *307* classified as *reliable* (*44.5%*) and *383* classified as *unreliable* (*55.5%*). Each instance is defined by *14* features and a binary class variable (*reliable* or *unreliable*).

Japanese Credit Screening (JCS): This dataset contains a number of people instances classified as *reliable* (granted credit) or *unreliable* (non granted credit). We used the numerical version, which is composed by *125* instances: *85* classified as *reliable* (*68.0%*) and *40* classified as *unreliable* (*32.0%*).

Each instance is defined by *16* features and a binary class variable (*reliable* or *unreliable*).

5.2 Strategy

All the experiments have been performed by adopting the *k-fold cross-validation* criterion, with $k=10$, in order to reduce the impact of data dependency, improving the value of the obtained results. In more detail, each dataset has been divided in k subsets, and each k subset has been used as test set, while the other $k-1$ subsets have been used as training set. The result is given by the average of all the obtained results.

We verified for the existence of a statistical difference between the results, by using the independent-samples *two-tailed Student's t-tests* ($p<0.05$).

Given that our approach needs a radius value ρ, it has been experimentally calculated for each dataset by testing a large range of values in the context of the training set I, choosing the value that leads towards the best performance in terms of *F-measure*. The results indicate as optimal ρ values *0.967* for the GCD dataset, *0.849* for the ACA dataset, and *0.789* for the JCS dataset, since these values maximize the *F-measure*.

5.3 Competitor

Although the literature indicates *Random Forest* as one of the most performing approach for the *Credit Scoring* [9], we have however carried out a preliminary study aimed to compare the AUC performance of ten binary classification approaches (i.e., *Naive Bayes*, *Logistic Regression*, *Multilayer Perceptron*, *Random Tree*, *Decision Tree*, *Logic Boost*, *SGD*, *Voted Perceptron*, *K-nearest*, and *Random Forests*), adopting the same cross-validation criterion used for the other experiments. The results indicate that *Random Forest* outperforms all the other approaches (GCD=*0.79*, ACA=*0.93*, JCS=*0.97*), so it is the only one we will be confronting with.

[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog

We have also performed a series of experiments in order to optimize the *Random Forest* performance. After we configured as *unlimited* the *maximum depth of the tree* parameter, we tested different values of the *number of randomly chosen attributes* (*nrca*) parameter. In order to avoid the overfitting problem, the tuning process has been performed by using both the training and testing sets and also in this case we adopted the same cross-validation criterion used for the other experiments. The results indicate as optimal *nrca* values *22* for the GCD dataset, *7* for the *ACA* dataset, and *15* for the JCS dataset.

5.4 Results

The analysis of the results shown in *Figure 2*, where, respectively, we have been compared the performance of our *Centroid Wavelet-based Approach* (CWA) to that of its competitor *Random Forests* (RF), in terms of *F-measure* and *Area Under ROC Curve* (AUC), leads towards the following considerations.

Figure 2.a shows that our CWA approach outperforms its competitor RF in terms of *F-measure* in the context of all the datasets, regardless of their size and their level of imbalance. This indicates its capability to classify the instances correctly, both with regard to the number of all performed classifications and to the number of the classifications that should have been made. Such result underlines two aspects, the first one related to the better performance achieved by it, while the second one related to the constancy of them. In fact, our approach outperforms RF in the context of all the datasets and its level of performance do not vary much (both in terms of quality and range), differently from its competitor.

Figure 2.b shows that our CWA approach reaches AUC performance similar (i.e., JCS dataset) or higher (i.e., GCD and ACA datasets) than that of its competitor RF, regardless of the size of data and the level of imbalance of them. The AUC metric measures the effectiveness of the evaluation model and the results indicate that our model gets higher performance than that of its competitor in the context of all datasets. It should be also noted how it obtains the best performance in a typical real-world scenario (i.e., the GCD dataset) characterized by many instances and a high degree of imbalance.

Additional considerations on the dimensionality reduction: In our DWT approach we have exploited only one of the two wavelet properties previously introduced (i.e., the multiresolution analysis) in order to mitigate the heterogeneity data issue through the pair-wise average and directed distances operations made by the Haar wavelet function. Now, we want to introduce the possibility to exploit the second properties (i.e., the dimensionality reduction) in order to reduce its computational complexity without a significant performance decay. Indeed, we can obtain a substantial curtailment of the processed elements (i.e., $|F|\cdot|I|$) by taking into account only the *pair-wise average* part of the *Haar wavelet function* output (small differences in values are to be attributed to the needed transformations to make $|F|=2^n$, with $n>1$.

The results show a reduction of *48.0%*, *46.0%*, and *47.0%* (respectively, in the GCD, ACA, and JCS dataset), without detecting any significant performance decay. It should be added that we get a similar result by using the *directed distances* part of the output instead of the *pair-wise average* one.

6. CONCLUSIONS AND FUTURE WORK

The *Credit Scoring* approach proposed in this paper is based on a threefold assessment of similarity carried out in the frequency-time domain offered by the *Discrete Wavelet Transformation*

process. The data analysis performed in this non-canonical domain through three metrics of similarity has led to a better capability to characterize the user instances in their correct class of destination (i.e., *reliable* or *unreliable*). Experimental results proved the effectiveness of such approach in the context of three datasets, where it outperforms its state-of-the-art competitor in typical real-world scenarios characterized by a considerable number of instances, regardless of their data distribution.

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 scenarios other than *Credit Scoring*.

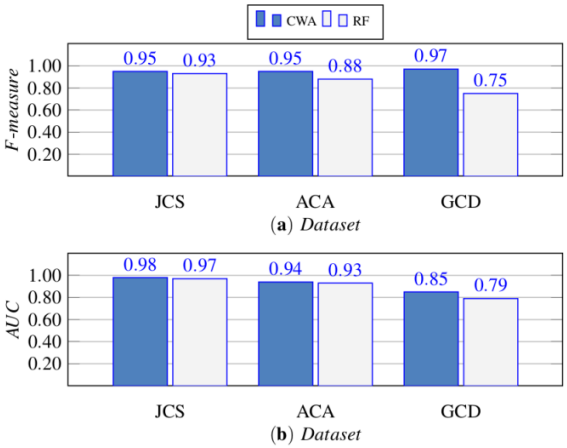

Figure 2. F-measure and AUC Performance

Figure 3.

7. ACKNOWLEDGMENTS

This research is partially funded by Regione Sardegna under project *Next generation Open Mobile Apps Development* (NOMAD), *Pacchetti Integrati di Agevolazione* (PIA) - Industria Artigianato e Servizi - Annualità 2013.

8. REFERENCES

[1] Abdou, Hussein A., and John Pointon. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent Systems in Accounting, Finance and Management* 18.2-3 (2011): 59-88.

[2] Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." *Applied Soft Computing* 6.2 (2006): 119-138.

[3] Attenberg, Josh, and Foster Provost. "Inactive learning?: difficulties employing active learning in practice." *ACM SIGKDD Explorations Newsletter* 12.2 (2011): 36-41.

[4] Baesens, Bart, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the operational research society* 54.6 (2003): 627-635.

[5] Chaovalit, Pimwadee, et al. "Discrete wavelet transform-based time series analysis and mining." *ACM Computing Surveys (CSUR)* 43.2 (2011): 6.

[6] Doumpos, Michael, and Constantin Zopounidis. "Credit scoring." *Multicriteria Analysis in Finance*. Springer International Publishing, 2014. 43-59.

Hand, David J. "Measuring classifier performance: a coherent alternative to the area under the ROC curve." *Machine learning* 77.1 (2009): 103-123.

[7] Japkowicz, Nathalie, and Shaju Stephen. "The class imbalance problem: A systematic study." *Intelligent data analysis* 6.5 (2002): 429-449.

[8] Lessmann, Stefan, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." *European Journal of Operational Research* 247.1 (2015): 124-136.

[9] Louzada, Francisco, Anderson Ara, and Guilherme B. Fernandes. "Classification methods applied to credit scoring: Systematic review and overall comparison." *Surveys in Operations Research and Management Science* (2016).

[10] Mallat, Stephane G. "A theory for multiresolution signal decomposition: the wavelet representation." *IEEE transactions on pattern analysis and machine intelligence* 11.7 (1989): 674-693.

[11] Percival, Donald B., and Andrew T. Walden. *Wavelet methods for time series analysis*. Vol. 4. Cambridge university press, 2006.

[12] Siami, Mohammad, and Zeynab Hajimohammadi. "Credit scoring in banks and financial institutions via data mining techniques: A literature review." *Journal of AI and Data Mining* 1.2 (2013): 119-129.