Addressing the EU Sovereign Ratings Using an Ordinal Regression Approach

Francisco Fernández-Navarro, Pilar Campoy-Muñoz, Mónica-de la Paz-Marín,
César Hervás-Martínez, *Member, IEEE*, and Xin Yao, *Fellow, IEEE*

***Abstract*—The current European debt crisis has drawn considerable attention to credit-rating agencies' news about sovereign ratings. From a technical point of view, credit rating constitutes a typical ordinal regression problem because credit-rating agencies generally present a scale of risk composed of several categories. This fact motivated the use of an ordinal regression approach to address the problem of sovereign credit rating in this paper. Therefore, the ranking of different classes will be taken into account for the design of the classifier. To do so, a novel model is introduced in order to replicate sovereign rating, based on the negative correlation learning framework. The methodology is fully described in this paper and applied to the classification of the 27 European countries' sovereign rating during the 2007–2010 period based on Standard and Poor's reports. The proposed technique seems to be competitive and robust enough to classify the sovereign ratings reported by this agency when compared with other existing well-known ordinal and nominal methods.**

***Index Terms*—Country risk detection, negative correlation learning (NCL), neural networks, ordinal regression.**

I. Introduction

NOWADAYS, credit-rating agencies (CRAs) play a crucial role in global financial markets due to the production of relevant credit information and its distribution to market participants, diminishing asymmetric information between investors and issuers. A wide range of stakeholders, including investors, issuers, and policy makers, use the rating agencies' information in their decision-making processes [1]. Moreover, Basel Capital Accord III has driven a renewed interest in credit ratings, allowing banks to employ them to determine the default probabilities of their debtors by calculating an adequate amount of capital to cover their credit risks (see, inter alia, [2] and [3]).

These CRAs have been in the spotlight during the ongoing European sovereign debt crisis. This crisis has been a theater of sovereign credit-rating downgrades, a widening of sovereign bond and credit default swap spreads and pressures on stock markets. Interestingly, financial markets throughout the Euro zone have been under pressure, although credit-rating actions concentrated on a few countries such as Greece, Iceland, Ireland, Portugal, and Spain [4]. The current debate about CRAs echoes previous discussions during the 1997–1998 Asian crisis [4], fuels the earliest criticisms about the correct assignment of ratings by these agencies [5], [6], and highlights the inherent conflicts of interest within their business model that is characterized by a lack of transparency and poor communication [7]. To overcome those limitations and promote competition among agencies, formal regulation of the credit-rating industry was introduced by the European Union in April 2009 and May 2011, and a proposal for a third regulation was published by the European Commission in November 2011. However, currently, the three largest CRAs, namely Moody's, S&P, and Fitch, continue dominating the sovereign ratings business, which is relatively new; S&P presented its first set of foreign currency sovereign ratings in January 1961 and Moody's followed in January 1974 [8].

In this regard, sovereign credit ratings could be considered a condensed assessment of a government's ability and willingness to repay its public debt both in principal and in interests in a timely fashion [9]. They are ordinal measures that can both reflect the current financial position of sovereign nations and provide an overview of their future financial positions. Thus, the relevance of sovereign credit ratings is reflected in several ways: 1) They improve the capability of countries to access international markets and to attract foreign investments [10]; 2) they face the growing international diversification of investors' portfolios and their needs for more accurate and greater amounts of information regarding country risk due to the impacts that a reassessment of a country's risk can have on its portfolios [11], [12]; 3) they can represent a ceiling for the ratings assigned to nonsovereign entities within a country [13], since there remains a "sovereign ceiling lite" [14], even though the sovereign ceiling rule has recently been eliminated by main agencies; 4) and, finally, the credit-rating events have a notable effect on financial

Manuscript received March 23, 2012; revised August 1, 2012 and December 7, 2012; accepted February 12, 2013. Date of current version November 18, 2013. This work was supported in part by the Spanish Inter-Ministerial Commission of Science and Technology under Project TIN2011-22794, the European Regional Development fund, and the "Junta de Andalucía" (Spain) under Project P2011-TIC-7508. The work of X. Yao was supported in part by the European Commission through FP7 Grant 270428. The work of F. Fernández-Navarro was supported by the "Junta de Andalucia" Predoctoral Program under Grant P08-TIC-3745. This paper was recommended by Associate Editor Prof. B. Chaib-draa, of the former Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews.

F. Fernández-Navarro is with the Department of Computer Science and Numerical Analysis, University of Córdoba, 14004 Córdoba, Spain, and also with the Advanced Concepts Team, European Space Research and Technology Centre, European Space Agency, 2201 AZ Noordwijk, The Netherlands (e-mail: i22fenaf@uco.es).

C. Hervás-Martínez is with the Department of Computer Science and Numerical Analysis, University of Córdoba, 14004 Córdoba, Spain (e-mail: chervas@uco.es).

P. Campoy-Muñoz and M.-de la Paz-Marín are with the Department of Economics, ETEA, University of Córdoba, 14004 Córdoba, Spain (e-mail: mpcampoy@gmail.com; ma2pamam@uco.es).

X. Yao is with the Centre of Excellence for Research in Computational Intelligence and Applications, School of Computer Science, University of Birmingham, Birmingham, B15 2TT, U.K. (e-mail: X.Yao@cs.bham.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2013.2247595

market developments, as negative news impact the country's bond and stock market [11], [15] and cause spillovers to other countries' equity and bond markets [16], while upgrades have limited or negligible effect [17], [18]. Rating agencies' signals also triggered foreign exchange market reactions [19].

To date, a wide range of classification techniques have been proposed in the literature on sovereign credit scoring. Recent empirical studies have emerged from two research lines. The first group focuses on the application of statistical techniques such as ordinary least squares [13], [20], [21], discriminant analysis (DA) [22], and ordered response models [23]–[25]. The second group makes use of artificial intelligence methods, especially machine-learning techniques such as neural networks [26], [27] or support vector machines (SVMs) [28]. However, recent studies suggested that the combination of multiple classifiers, which is known as ensemble learning, may lead to better performance [29], [30].

In this paper, the sovereign rating problem has been addressed using an ordinal regression approach because of the ordinal nature of the dependent variable. Ordinal regression algorithms are intended to take advantage of this order information in order to improve the classification performance. An ordinal categorical variable may be inherently a continuous variable termed as a grouped continuous variable [31]. On the other hand, an ordinal categorical variable such as attitude or opinion can be considered a manifestation of an underlying continuous variable (see, e.g., [32]). In both cases, it is sensible to relate the observed ordinal categorical variable to an underlying continuous variable.

A more commonly used idea is to transform the ordinal regression problem into a multinomial classification one, or to add additional constraints to traditional classification formations. In this sense, Crammer and Singer [33] generalized the online perceptron algorithm with multiple thresholds in order to seek the direction and thresholds for ordinal regression. Shashua and Levin [34] proposed two large-margin principles, namely, a fixed-margin principle and a sum of margins principle, to handle direction and multiple thresholds. More recently, Chu, together with Ghahramani, presented a probabilistic kernel approach to ordinal regression based on Gaussian processes [35] and, together with Keerthi, two new SVM approaches for ordinal regression, which optimized multiple thresholds to define parallel discriminant hyperplanes for ordinal scales [36]. In [37], a novel framework for ordered classes, based on replicating the dataset, was introduced in the context of SVMs. Finally, Sun *et al.* [38] have proposed a new support vector approach for ordinal regression which maximizes the sum of the margins for the computation of parallel discriminant hyperplanes because they argue that, although SVM-based methods have shown great promise in ordinal regression, they do have some drawbacks.

On the other hand, an ensemble of multiple classifiers is expected to reduce generalization error by considering the opinions from multiple classifiers. Therefore, diversity becomes an important issue to take into account [39], [40]. If every classifier had the same opinion, then the construction of multiple classifiers would become meaningless. Diversity can be described as the degree to which classifiers come to different decisions about the same problem. This degree allows voted accuracy to be greater than the one achieved by a single classifier.

Concerning ordinal regression problems, there are several ensemble-related approaches in the literature. These approaches transform the ordinal classification problem into a nested binary classification one and then combine the resulting classifier predictions to obtain the final decision. For instance, Frank and Hall [41] proposed a general algorithm that enables standard binary classifiers to make use of order information in attributes. Waegeman and Boullart [42] imposed explicit weights over the patterns of each binary classifier, in such a way that errors on training patterns are penalized proportionally to the absolute difference between their rank and its ranking. However, both approaches take a base binary classification algorithm rather than a base ordinal regression one.

Negative correlation learning (NCL) [43], [44] is an ensemble learning method of neural networks that uses different mechanisms to induce diversity. It has shown a number of empirical successes in various applications and competitive results. NCL tries to induce diversity directly by changing its error function into a "diversity-encouraging" one. In this study, the NCL methodology was adapted to ordinal regression problems and validated using an EU sovereign-rating dataset. To the best of the authors' knowledge, this research work would be the first adaptation of the NCL to ordinal regression problems.

The aim of this paper is to ascertain whether the NCL methodology can improve the classification accuracy when compared with the more common methods employed in machine learning. To achieve this aim, this study carries out an empirical analysis of foreign currency sovereign-debt ratings with the data on 27 EU member countries provided by S&P during the financial crisis between 2007 and 2010. The S&P foreign currency long-term rating history is selected instead of other agencies' historical ratings due to data availability. Moreover, S&P seems more prone to produce more rating revisions and to lead the rerating of other agencies [18], [45]. Foreign currency rating announcements by S&P also seem to convey a greater stock market impact on the country itself and do not seem to be fully expected by the market [11]

The remainder of this paper proceeds as follows. The next section reviews the relevant literature on classification methodologies in rating assignments. Second, the methodology applied in this paper and its motivation are described; third, the experimental study is carried out with the description of the variables and the database employed, and the presentation of empirical results, the discussion and, finally, the main conclusions about these experimental results.

II. Literature Review

Although a number of methodologies have been employed in the empirical modeling of credit ratings, the main focus was on corporate bonds rather than on sovereign risk [26]. In the latter case, the earliest approach applies statistical techniques such as linear regression methods [13], [20], [21], ordered response models [23]–[25], and DA, which are derived from previous works of Frank and Cline and Sargen [22], [46], or Carleton and Lerner [47], which refer to subsovereign entities.

While logit, probit, and DA models require such assumptions as the normality of variable distribution and independent predictors, nonparametric and nonlinear models such as artificial neural networks (ANNs) [48]–[50] do not rely on these assumptions that are adopted to turn traditional statistical methods into more tractable ones. Comparing ANNs and statistical models for sovereign risk prediction, most related studies pointed out that ANNs outperformed statistical models. Bennell *et al.* [26] demonstrate that ANNs represent improved technology to calibrate and predict sovereign ratings compared with ordered probit modeling. In this regard, Yim and Mitchell [27] also prove that hybrid ANNs produce better results than ordered response models and DA. However, some authors pointed out that ANNs have difficulty in determining the difference between adjacent rating classes [51]. In fact, Huang *et al.* [52] found out that the probabilities of misclassification within one class away from the real one were over 90%. Since ANNs do not require the prior specification of the theoretical model, it is argued that their functional form is totally unrelated to the economic theory [53]. ANNs are constructed from the data and not from an economic theory. However, when economic theory presents difficulties to be successfully implemented, the use of ANN models can be an acceptable option to have good predictions on the EU sovereign ratings (while the economic community researches new alternatives). ANNs also have other technical drawbacks including the risk of overfitting, difficulty in determining the values of control parameters, and the number of processing elements in the hidden nodes [28].

The SVM algorithm has recently become a solution often employed to solve prediction problems because of its robustness and high accuracy. An SVM solution is globally optimal because SVMs seek to minimize structural risk. Conversely, the solutions found by ANN models tend to fall into local optimum because they seek to minimize empirical risk. SVMs were originally designed for two-class tasks and, therefore, are not naturally geared for multiclass classifications, which apply to credit ratings: even so researchers have made attempts to extend the original SVM to multiclass classification (MSVM). The prior studies that applied MSVMs to credit-rating issues were not designed to reflect the ordinal nature of this domain (see, e.g., [52]), although recently there have been developments which take ordinal characteristics into account to efficiently and effectively handle multiple ordinal classes such as in [28]. In this study, the MSVM approach improves the performance of classification as compared with other typical multiclass classification techniques. However, Huang *et al.* [52] pointed out that there are two major aspects that should be considered cautiously when SVM methods are applied to solve rating problems. The first is how to select the optimal input feature, and the second is how to set good kernel parameters.

As can be seen, both statistical and machine-learning techniques have been explored for credit rating, although there are no consistent conclusions about which performs better. Accordingly, most recent studies suggest combining multiple classifiers, i.e., ensemble learning, which have shown better performance than any individual methods. Ensemble models, in general, have been successfully used for classification and regression. They not only introduce more stable predictions through linear combination, but also provide sufficient power to approximate complicated target functions [54]. Ensembles are trending in financial topics such as the prediction of financial crisis or the credit scoring due to these techniques outperform single classification techniques, but to the best of our knowledge, this technique has not been applied to sovereign credit-rating problems [55].

The next section introduces the motivation for using ensembles in ordinal regression problems, and then, the proposed ensemble methodology is described.

III. Ensemble Learning Suitability for Ordinal Regression Problems

The aim of this paper is to validate ensemble models in the ordinal regression problem of classifying the sovereign ratings provided by the S&P international agency. Therefore, this section describes the motivation for the use of ensemble models in ordinal regression problems.

In a regression problem, we can define why and how differences between individual predictor outputs contribute to overall ensemble accuracy. However, in a classification problem, there is no such neat theory. There is a subtle point here, often overlooked. Difficulties in quantifying classification error diversity are not intrinsic to ensemble-tackling classification problems. It is possible to reformulate any classification problem as a regression one by approximating posterior class probabilities [56]. For the regression context, the question can be clearly phrased as "how could we quantify diversity when our predictors are output real-valued numbers and are combined by a convex combination?" For the case of Tumer and Ghosh [56] study, the question is the same, except for the fact that the "real-valued" numbers are probabilities. A much harder question appears when we are restricted in that our predictors can only output discrete class labels. In this case, the outputs do not present any intrinsic order among them, and therefore, the concept of "covariance" is not so simple.

A step toward understanding this question can be taken by considering where the bias-variance-covariance decomposition comes from: It falls neatly out of the bias-variance decomposition of the ensemble error. However, when our classification of a pattern is either correct or incorrect, we have a zero-one loss function.

Tumer and Ghosh [56], [57] provided a theoretical framework to analyze the simple averaging combination rule when our predictor outputs are estimates of the posterior probabilities of each class.

For an input variable of only one dimension, x, the solid curves show the true posterior probabilities of $\mathcal{C}_1$ and $\mathcal{C}_2$, these are $P(\mathcal{C}_1)$ and $P(\mathcal{C}_2)$, respectively (see Fig. 1). The dotted curves show estimates of the posterior probabilities, from one of the individuals of the ensemble, these are $\hat{P}(\mathcal{C}_1)$ and $\hat{P}(\mathcal{C}_2)$. The solid vertical line at θ indicates the optimal decision boundary. The dark shaded area, which is called the Bayes error, is an

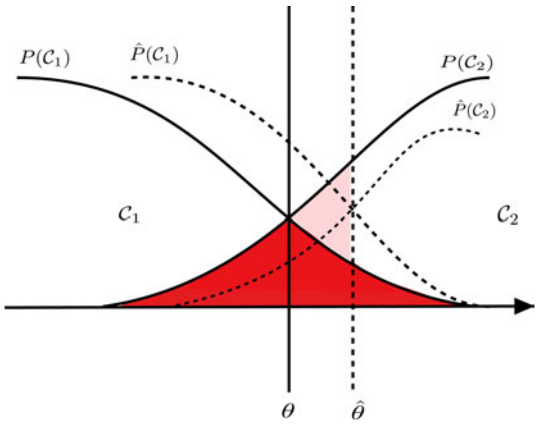


Fig. 1. Tumer and Ghosh's analysis of ordinal regression error.

irreducible quantity. The dotted vertical line at $\hat{\theta}$ indicates the boundary placed by the individual, which is a certain distance from the optimal. The light shaded area indicates the added error that the predictor makes in addition to the Bayes error. The individual i approximates the posterior probability of $\mathcal{C}_1$ as

$$\hat{P}_i(\mathcal{C}_1|x) = P_i(\mathcal{C}_1|x) + \varepsilon_i(\mathcal{C}_1|x) \tag{1}$$

where $P_i(\mathcal{C}_1|x)$ is the true posterior probability of $\mathcal{C}_1$, and $\varepsilon_i(\mathcal{C}_1|x)$ is the estimation error. Let us assume the estimation errors on different classes $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent and identically distributed random variables with zero mean and variance σ_{ε_i}. The individual's expected added error of classifying classes $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as

$$E_{\mathrm{add},i} = \frac{2\sigma^2_{\varepsilon_i}}{P_i^{'}(\mathcal{C}_1|x) - P_i^{'}(\mathcal{C}_2|x)} \tag{2}$$

where $P_i^{'}(\mathcal{C}_1|x)$ and $P_i^{'}(\mathcal{C}_2|x)$ are the derivatives of the true posterior probability of classes $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. If the decision boundary was determined by an ensemble of individuals, the authors show that the expected added error of the ensemble is

$$E_{\mathrm{add}}^{\mathrm{ens}} = E_{\mathrm{add}} \left(\frac{1 + \delta(M-1)}{M} \right) \tag{3}$$

where M is the number of classifiers, E_{add} is the expected added error of the individual classifiers (they are assumed to have the same error), and δ is a correlation coefficient. If δ is zero (the classifiers in the ensemble are statistically independent because they are normally distributed), we have $E_{\mathrm{add}}^{\mathrm{ens}} = \frac{1}{M} E_{\mathrm{add}}$, i.e., the error of the ensemble will be M times smaller than the error of the individuals. If δ is 1, i.e., perfect correlation, then the error of the ensemble will just be equal to the errors of the individuals ($E_{\mathrm{add}}^{\mathrm{ens}} = E_{\mathrm{add}}$).

Taking the work of Tumer and Ghosh into consideration, it seems reasonable to think that in the case of ordinal regression, we should take the correlation between individuals into account. However, to our knowledge, there is no previous work where individuals were designed taking into account the correlation among them.

```
Threshold Regression Algorithm:
Estimate a potential function $f(\mathbf{x})$ that predicts (a monotonic transform of) the real-valued outcomes.
Determine a threshold vector $\theta \in \mathbb{R}^{J-1}$ to represent the intervals in the range of $f(\mathbf{x})$, where $\theta_1 \le \theta_2 \le \ldots \le \theta_{J-1}$
```

Fig. 2. Threshold regression algorithm.

IV. Description of the Negative Correlation Methodology in Ordinal Regression

Ordinal regression is similar to regression because the labels in both $\mathcal{Y}_r$ and $\mathbb{R}$ represent ordinal information. Nevertheless, unlike the real-valued regression labels in $\mathbb{R}$, the discrete ranks in $\mathcal{Y}_r$ do not carry metric information. That is, ordinal ranking deals with qualitative ranks, while regression focuses on quantitative, real-valued outcomes. To model ordinal ranking problems from a regression perspective, it is often assumed that some underlying real-valued outcomes exist, although they are unobservable [31]. The hidden local scales "around" different ranks can be quite different, but the actual scale (metric) information is not encoded in the ranks.

Under the above assumption, each rank represents a contiguous interval on the real line. Then, ordinal ranking can be approached by the algorithm described in Fig. 2.

In the threshold regression algorithm, the potential function tries to uncover the nature of the assumed underlying outcome, and the threshold vector estimates the possibly different scales around different ranks. The two abstract steps of the algorithm are indeed taken by many existing ordinal ranking algorithms. For instance, in the GPOR algorithm of Chu and Ghahramani [35], the potential function $f(\mathbf{x})$ is assumed to follow a Gaussian process, and the threshold vector $\boldsymbol{\theta}$ is determined by Bayesian inference with respect to some noise distribution. In the PRank algorithm of Crammer and Singer [33], the potential function $f(\mathbf{x})$ is taken to be a linear function and the pair $\langle f(\mathbf{x}), \boldsymbol{\theta} \rangle$ are updated simultaneously. Some other algorithms are based on SVM, and they work on potential functions of the form $f_v(\mathbf{x}) = \langle v, \phi(\mathbf{x}) \rangle$, where $\phi(\mathbf{x})$ maps $\mathbf{x} \in \mathbb{R}^k$ to some Hilbert space [36].

In our proposal, the thresholds are fixed *a priori*, and they are not modified during the training procedure. Therefore, each threshold is defined as $\theta_j = \theta_1 + (j-1)\vartheta$, where ϑ represents the width of the intervals, $j = 2, \ldots, J-1$, and $\theta_1 \in \mathbb{R}$ (in this approach, the parameter ϑ takes a constant value for simplicity). That is due to the fact that to measure the diversity in the proposed model, it becomes necessary that all individuals in the ensemble project the patterns using the same thresholds setting (to measure the diversity in a common space). Furthermore, a nonlinear model was applied (ensemble of ANNs). Considering this kind of models, it is relatively easy to converge toward an acceptable solution without modifying the thresholds. In addition, an optimal projection generated with adaptive thresholds can be transformed into an equivalent one with fixed thresholds by linearly scaling the patterns class by class.

As previously stated, a standard multilayer perceptron (MLP) (nonlinear ranking function) is considered as the potential

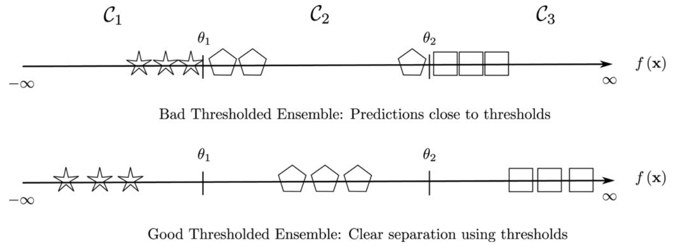


Fig. 3. Example of good and bad thresholded ensemble.

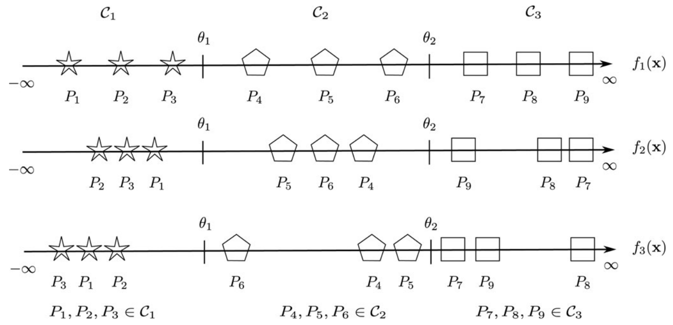

$\mathcal{C}_1$
$\mathcal{C}_2$
$\mathcal{C}_3$
θ_1
θ_2
$f_1(\mathbf{x})$
$-\infty$
∞
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9
θ_1
θ_2
$f_2(\mathbf{x})$
$-\infty$
∞
P_2 P_3 P_1 P_5 P_6 P_4 P_9 P_8 P_7
θ_1
θ_2
$f_3(\mathbf{x})$
$-\infty$
∞
P_3 P_1 P_2 P_6 P_4 P_5 P_7 P_9 P_8
$P_1, P_2, P_3 \in \mathcal{C}_1$
$P_4, P_5, P_6 \in \mathcal{C}_2$
$P_7, P_8, P_9 \in \mathcal{C}_3$

Fig. 4. Example of three different ranking functions: All of these ranking functions are diverse and accurate.

function $f(\mathbf{x})$ of each individual. Finally, the potential ensemble function is determined by simple averaging

$$\overline{f}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\mathbf{x}) \tag{4}$$

where $\overline{f}(\mathbf{x})$ is the average output of the whole ensemble of M networks. This way, the threshold ensemble model determines the class of each pattern as

$$r_{(\overline{f}(\mathbf{x}),\theta)} = \min\{k : \overline{f}(\mathbf{x}) \le \theta_k\} = \max\{k : \overline{f}(\mathbf{x}) > \theta_{k-1}\}. \tag{5}$$

It is well known that a multicriteria search for an ensemble that maximizes both accuracy and diversity leads to more accurate ensembles than optimizing a single criterion. Intuitively, we expect the potential value $f(\mathbf{x})$ to be in the desired interval $(\theta_{i-1}, \theta_i]$, and we want $f(\mathbf{x})$ to be far from the boundaries (thresholds).

Fig. 3 shows two threshold ensemble models. In the first case, the predictions are very close to the thresholds. In this case, if there is noise in the data, it could cause changes in the predictions. In the second case, the predictions are as far as possible from the thresholds. In this case, if there is noise in the data, the changes in predictions would be lower.

However, in ensemble modeling, ensemble diversity is another key aspect. Fig. 4 shows three threshold ensemble models. All three models are perfectly accurate. However, as noted at the beginning of this section, these models assume that there is a latent variable which is defined on the real line where the patterns are projected. Taking this into account, which one of the three projections best represents the real projection of the patterns? It is impossible to answer this question because we do not know the values for each of the patterns in the latent variable. If these values had been known, we would have considered the problem to be a standard regression problem.

In our opinion, it would be very interesting to combine individuals in the ensemble so that they meet two objectives: First, their projections must be as far as possible from the thresholds, and second, their projections must be as different as possible from the projections of the remaining individuals in the ensemble. With this second objective, we ensure that the individuals in the ensemble are accurate, even though their projections are different from the projections of other individuals. With the average of the projections, our aim is to better estimate the real values of the latent variable. This paper optimizes both objectives using the NCL framework [44], [58].

NCL uses the following regularization term to determine the amount of correlation in the ensemble [59], [60]:

$$\begin{aligned} R = p_i &= \frac{1}{N}\sum_{n=1}^{N}\left(f_i(\mathbf{x}_n) - \overline{f}(\mathbf{x}_n)\right)\left(\sum_{j\neq i} f_j(\mathbf{x}_n) - \overline{f}(\mathbf{x}_n)\right)\\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(f_i(\mathbf{x}_n) - \overline{f}(\mathbf{x}_n)\right)^2; i = 1,\ldots,M. \end{aligned}$$

Therefore, using this approach, the error of neural network i becomes

$$e_i = \frac{1}{N}\sum_{n=1}^{N}(f_i(\mathbf{x}_n) - y_n)^2 + \lambda R; i = 1,\ldots,M \tag{6}$$

where λ is a weighting parameter on the regularization term R, y_n represent the rank of the nth pattern, and N is the number of patterns. The λ parameter controls a tradeoff between the two terms; with $\lambda = 0$, we would have an ensemble with each network trained with backpropagation, and as λ increases, more and more emphasis would be placed on minimizing the regularization term.

In this study, the model parameters are optimized using the $iRprop^+$ local improvement procedure. In this case, the gradient vector is given by the following equation:

$$\nabla e_i = \left(\frac{\partial e}{\partial \beta_1}, \ldots, \frac{\partial e}{\partial \beta_S}, \frac{\partial e}{\partial \mathbf{w}_1}, \ldots, \frac{\partial e}{\partial \mathbf{w}_S}\right) \tag{7}$$

where $\boldsymbol{\beta}$ are the connections between hidden and output layers, and $\boldsymbol{w}$ are the connections between input and hidden layers of the MLP neural network model $f_i(\mathbf{x}_n)$.

Let η be any of the parameters of $\boldsymbol{\beta}$ or $\boldsymbol{w}$. Therefore

$$\begin{aligned} \frac{\partial e_i}{\partial \eta} = &\left(\frac{2}{N}\sum_{n=1}^{N}(f_i(\mathbf{x}_n) - y_n)\frac{\partial f_i(\mathbf{x}_n)}{\partial \eta}\right) - \\ &\left(\frac{2\lambda}{N}\left(1 - \frac{1}{M}\right)\sum_{n=1}^{N}(f_i(\mathbf{x}_n) - \overline{f}(\mathbf{x}_n))\frac{\partial f_i(\mathbf{x}_n)}{\partial \eta}\right). \end{aligned}$$

The following expressions include the derivatives of the parameters for the f_i MLP model ($f_i = \sum_{s=1}^{S} \beta_s B_s(\mathbf{x}, \mathbf{w}_s)$):

$$\frac{\partial f_i}{\partial \beta_s} = B_s(\mathbf{x}, \mathbf{w}_s);\ 1 \le s \le S.$$

The gradient for the hidden layer depends of the kind of the basis function. In this paper, $B_s(\mathbf{x}, \mathbf{w}_s)$ are sigmoidal nodes:

$$B_s(\mathbf{x}, \mathbf{w}_s) = \sigma\left(\sum_{i=1}^{k} w_{is} x_i\right), \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{\partial f_i}{\partial \mathbf{w}_s} = \left(\frac{\partial f_i}{\partial w_{s1}}, \ldots, \frac{\partial f_i}{\partial w_{sk}}\right)$$

$$\frac{\partial f_i}{\partial w_{st}} = \beta_s \sigma' \left(\sum_{i=1}^{k} w_{is} x_i\right) x_t$$

where $s = 1, 2, \ldots, S$ and $t = 1, 2, \ldots, k$.

V. Computational Experiments and Results

A. Data and Variables Involved in Country Risk Detection

The analysis covers 27 EU sovereign borrowers during the period from 2007 to 2010, encompassing the worldwide economic downturn and the beginnings of the on-going EU debt crisis: Austria, Belgium, Bulgaria, Czech Republic, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and the U.K.

The output variable for the NCL is the long-term foreign currency rating assigned to each sovereign state by S&P included in the analysis as of December 31 of each year. The S&P's scale of measurement is a 22-point risk-rating scale, which is replaced by a numerical equivalent grade into broad categories maintaining the ordinal ranking of creditworthiness [21] (see Table I). In the case under study, the S&P ratings of EU countries are located on the top five broadest categories (see Table II). As shown in Table III, the number of sovereign states having their credit rating downgraded has outnumbered the number of those that have had their rating upgraded, leading some to conclude that the credit quality of the EU sovereign debt has declined over time. However, empirical studies have emphasized that CRAs have become more conservative over time, tightening the requirements that issuers must fulfill in order to achieve higher credit ratings [61], [62]. Regarding sovereign issuers, Ferri *et al.* [63] found that given their economic fundamentals, some countries were downgraded excessively during the East Asian financial crisis. In the same way, Gartner *et al.* [64] pointed out that the rating for the so-called PIGS countries (Portugal, Ireland, Greece, and Spain) during the ongoing EU debt crisis is 2.30 notches lower than that of a hypothetical country, which has the same economic fundamentals but does not belong to this group. It is important to note that the dependent variable is the credit rating assigned to the country by analysts within S&P. Therefore, the model in this paper attempts to replicate the S&P sovereign ratings, rather than reproducing the decision process undertaken by the analyst or providing a more accurate estimate of sovereign rating.

TABLE I
S&P Credit-Rating Measures

S&P		Rating
Highest quality	AAA	$\mathcal{C}_1$
High quality	AA+	$\mathcal{C}_2$
	AA	$\mathcal{C}_2$
	AA-	$\mathcal{C}_2$
Strong payment capacity	A+	$\mathcal{C}_3$
	A	$\mathcal{C}_3$
	A-	$\mathcal{C}_3$
Adequate payment capacity	BBB+	$\mathcal{C}_4$
	BBB	$\mathcal{C}_4$
	BBB-	$\mathcal{C}_4$
Likely to fulfil obligations	BB+	$\mathcal{C}_5$
	BB	$\mathcal{C}_5$
	BB-	$\mathcal{C}_5$
High Credit Risk	B+	$\mathcal{C}_6$
	B	$\mathcal{C}_6$
	B-	$\mathcal{C}_6$
Very High Credit Risk	CCC+	$\mathcal{C}_7$
	CCC	$\mathcal{C}_7$
	CCC-	$\mathcal{C}_7$
Near default with possibility of recovery	CC	$\mathcal{C}_8$
Default	SD	$\mathcal{C}_9$
	D	$\mathcal{C}_9$

TABLE II
S&P Ratings, 2007–2010

S&P		Ratings			
Rating	Class	2007	2008	2009	2010
AAA	$\mathcal{C}_1$	11	11	9	9
AA	$\mathcal{C}_2$	3	3	4	3
A	$\mathcal{C}_3$	9	8	8	9
BBB	$\mathcal{C}_4$	4	5	4	3
BB	$\mathcal{C}_5$	0	0	2	3

TABLE III
S&P Rating Movements, 2007–2010

	Rating movements		
	2007-2008	2008-2009	2009-2010
Upgrades	2	0	2
Downgrades	5	8	4
	Rating movements in rating notches		
Upgrades	1.0	0.0	1.0
Downgrades	1.2	1.4	2.0

Rating agencies, due to their business practice, do not officially disclose the precise models used for their rating methodologies. A common practice among rating agencies is to assign qualitative scores to several criteria and then arrive at a weighted average score. Beers and Cavanaugh [65] provide an excellent explanation of the criteria used by Standard and Poor's. They list 44 variables grouped into ten categories—political risk, income and economic structure, economic growth prospects, fiscal flexibility, general government debt burden, off-budget and contingent liabilities, monetary flexibility, external liquidity, and public sector external debt burden. Nevertheless, many researchers have found that the ratings by major agencies are largely explained by a handful of macroeconomic variables (see [66]).

To our knowledge, Feder and Uy [67] were the first to identify the determinants of country risk ratings. Many other studies investigate the determinants of sovereign ratings and show that sovereign ratings are mainly driven by economic fundamentals (see, e.g., [13] and [20]). Therefore, in line with the theoretical and empirical literature, ten economic indicators, as well as one political indicator, have been selected as explanatory variables.

1) *Real GDP growth*: Indicator of the country's government's ability to repay outstanding obligations. Unit of measurement: Rate (Eurostat).
2) *GDP per capita*: Targets total income of country's citizen—reflects cost of living. Unit of measurement: Euros per inhabitant (Eurostat).
3) *Goverment debt*: Indicates the total debt of government held by the public. Unit of measurement: Percent of GDP (Eurostat).
4) *Fiscal balance*: Indicates the demand (deficit)/offer (surplus) of external budgetary financing. Unit of measurement: Percent of GDP (Eurostat).
5) *External debt*: Indicates the outstanding amount of those current, and not contingent liabilities owed to nonresidents by residents of an economy that requires payment(s) either of principal and/or interest by the debtor at some point(s) in the future. Unit of measurement: Percent of exports (World Bank and EU Member Central Banks).
6) *Level of external reserves*: Held by a country's central bank as defense against withdrawal of foreign credit. Unit of measurement: Percent of imports (Eurostat).
7) *Current account balance*: Indicate how much net import of capital a country requires. Unit of measurement: Percent of GDP (Eurostat).
8) *Inflation*: Shows change in the level of price index for a basket of commonly used goods. Unit of measurement: Index (2005=100) (Eurostat).
9) *Unemployment rate*: Indicator of size of output gap and of underutilization of resources. Unit of measurement: Rate (Eurostat).
10) *Unit labor cost*: Indicator of a country's competitiveness in international trade. Unit of measurement: Index (2005=100) (Eurostat).
11) *Government effectiveness*: Indicator of the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Unit of measurement: Percentile (World Bank)

The relationship between these variables and credit ratings is such that it is expected that higher levels of external and government debt, higher rates of inflation and unemployment, and higher labor costs, and current account deficits will result in lower ratings; on the other hand, higher credit ratings are associated with higher levels of fiscal balance (surplus), higher levels of income and external reserves, higher rates of GDP growth, and, finally, higher scores in government effectiveness. Furthermore, the variables of inflation, unemployment, GDP growth, fiscal balance, and current accounts are entered as a three-year average, reflecting the agencies' approach to eliminating the effect of the business cycle when deciding on a sovereign rating. Recently, during 2009 and 2010, CRAs have started to define new methodologies that shift the criteria from a "through-a-cycle" to a "through-a-crisis" approach. In particular, S&P has established a new credit stability criterion that uses hypothetical stress scenarios as benchmarks for calibrating the criteria over time. However, neither those scenarios nor other further developments are considered in our experimental design, as they will not be effective until the second semester of 2011.

In order to assess the ability of the models in an out-of-time dataset, the dataset has been split into two subsequent time periods, holding the later for evaluation of the model only. Consequently, the training information table consisted of 81 samples from 27 countries in the period 2007–2009 (annual data) described by the variables explained. The test or generalization information table consisted of 27 data described by the same variables in the year 2010.

B. Machine-Learning Methods Used for Comparison Purposes and Experimental Design

For comparison purposes, different state-of-the-art methods have been included in the experimentation. These methods are the following.

1) *Nominal Classifiers:*
   a) The *Multilogistic Regression (MLR)* algorithm. It is based on applying the LogitBoost algorithm with simple regression functions and determining the optimum number of iterations by a fivefold cross validation [68].
   b) A *Gaussian Radial Basis Function Network (RBFN)* [69], deriving the centers and width of hidden units using k-means, and combining the outputs obtained from the hidden layer using logistic regression.
   c) An *MLP* [69] with sigmoid units as hidden nodes, obtained by means of the back-propagation algorithm.
   d) An *SVM* [70] nominal classifier is included in the experiments in order to validate our proposal contributions. Cost support vector classification (SVC) available in libSVM 3.0 [71] is used as the SVM classifier implementation.
2) *Ordinal Regression approaches:*
   a) The *Proportional Odd Model (POM)* [31] is an extension of the binary logistic regression model for ordinal multiclass categorization problems. This is one of the first models specifically designed for ordinal regression, and it arose from a statistical background. The model is based on the assumption of stochastic ordering of the space X, i.e., for all $\mathbf{x}_1$ and $\mathbf{x}_2$, too:

$$P(y \preceq \mathcal{C}_j|\mathbf{x}_1) \succeq P(y \preceq \mathcal{C}_j|\mathbf{x}_2) \;\; \forall\, \mathcal{C}_j \in Y$$

or

$$P(y \preceq \mathcal{C}_j|\mathbf{x}_1) \preceq P(y \preceq \mathcal{C}_j|\mathbf{x}_2) \;\; \forall\, \mathcal{C}_j \in Y. \quad (8)$$

In the POM model, the cumulative probability of the rating y is given by

$$P(y \preceq \mathcal{C}_j) = \frac{1}{1 + \exp(\theta_j + \beta_1 x_1 + \cdots + \beta_k x_k)} \tag{9}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ represents the array with k inputs variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ the corresponding coefficient vector, and $j = 1, \ldots, J$. Because $P(y \leq \mathcal{C}_J) = 1$, the parameter θ_J is equal to ∞. The latent variable is the linear combination of the input variables

$$z = -\beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k \tag{10}$$

and summarizes the financial information of the risk entity.

b) *SVMRank* [72] applies the extended binary classification (EBC) method to SVM. The EBC method can be summarized in the following three steps. First, all training samples are transformed into extended samples, weighting these samples by using the absolute cost matrix. Second, all the extended examples are jointly learned by a binary classifier with confidence outputs, aiming at a low weighted $0/1$ loss. The last step is used to convert the binary outputs to a rank.

c) *Support Vector Ordinal Regression (SVOR)* by Chu and Keerthi [73], [74] proposes two new support vector approaches for ordinal regression. Here, multiple thresholds are optimized in order to define parallel discriminant hyperplanes for the ordinal scales. The first approach with explicit inequality constraints on the thresholds derives the optimality conditions for the dual problem, and adapts the SMO algorithm for the solution: We will refer to it as SVOR(EX). In the second approach, the samples in all the categories are allowed to contribute errors to each threshold, which is why there is no need to include inequality constraints in the problem. This approach is called an SVOR with implicit constraints (SVOR(IM)).

d) *Ensemble Approaches for Ordinal Regression:*

   i) *A Simple Approach to Ordinal Regression (ASAOR)* [41] is a metaclassifier that allows standard classification algorithms to be applied to ordinal class problems. This methodology was already described in Section I. In this study, the C4.5 method that is available in Weka [75] is used as the underlying classification algorithm, since this is the one initially employed by the authors of ASAOR.

   ii) *Multiclass Ordinal Support vector machines (MCOSvm)* [42] is an enhanced ensemble method for ordinal regression. It is closely related to the ASAOR method. In this case, weighted SVMs are used as base classifiers. Specific weights are assigned to each pattern in such a way that errors of more than one rank are more penalized. Therefore, the weight of a training pattern differs for each binary SVM.

3) *Regression approaches:*

   a) *Regression Neural Network Model (rNN)*: As stated in Section I, regression models can be applied to solve the classification of ordinal data. A common technique for ordered classes is to estimate by regression any ordered scores $s_1 < \cdots < s_J$ by replacing the target class $\mathcal{C}_i$ with the score s_i. The simplest case would be setting $s_i = i$; $i = 1, \ldots, J$. A neural network with a single output was trained to estimate the scores. Furthermore, this model is particularly interesting because our ensemble model is composed of M rNN-type individuals.

Regarding the hyperparameters of different algorithms, the following procedure has been applied. For the support vector algorithms, i.e., SVC, SVMRank, SVOR(EX), SVOR(IM), and MCOSvm, the corresponding hyperparameters (regularization parameter, C, and width of the Gaussian functions, γ) were adjusted using a grid search with a fivefold cross validation, considering the following ranges: $C \in \{10^3, 10^1, \ldots, 10^{-3}\}$ and $\gamma \in \{10^3, 10^0, \ldots, 10^{-3}\}$.

For the neural network algorithms, i.e., MLP and rNN, the corresponding hyperparameters (number of hidden neuron, H, and number of iterations of the local search procedure, iterations) were adjusted using a grid search with a fivefold cross validation, considering the following ranges: $H \in \{5, 10, 15, 20, 30, 40\}$ and iterations $\in \{25, 50, \ldots, 500\}$. In the case of the RBFN methodology, the number of hidden neuron, H, was adjusted using a grid search with a fivefold cross validation, considering the following ranges: $H \in \{5, 10, 15, 20, 30, 40\}$. Finally, the NCL selects the λ parameter by cross validation within the range $\{0.0, 0.1, \ldots, 1.0\}$. We notice that λ could be a little greater than 1 $[\lambda \leq \frac{M}{M-1}]$ to guarantee the positive definiteness of the Hessian matrix [59]. Since we use $M = 25$ in this paper, the upbound of λ (1.0417) is close to 1 and we will not use the λ values which are greater than 1.

Regarding the experimental design, the rNN and NCL approaches are nondeterministic methodologies because the neural network weight vectors are initialized randomly. For these methodologies, we run the procedure 30 times for the holdout considered.

C. Ordinal Classification Evaluation Metrics

Four evaluation metrics have been considered which quantify the accuracy of N predicted ordinal labels for a given dataset $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$, with respect to the true targets $\{y_1, y_2, \ldots, y_N\}$.

1) Correct classification rate (C) is simply the fraction of correct predictions on individual samples:

$$C = \frac{1}{N} \sum_{i=1}^{N} I\left(\hat{y}_i = y_i\right) \tag{11}$$

where $I(\cdot)$ is the zero-one loss function, and N is the number of patterns of the dataset.

2) Mean absolute error (MAE) is the average deviation of the prediction from the true targets, i.e.,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\mathcal{O}(\hat{y}_i) - \mathcal{O}(y_i)| \tag{12}$$

where $\mathcal{O}(\mathcal{C}_k) = k, 1 \leq k \leq K$, i.e., $\mathcal{O}(y_i)$ is the order of class label y_i.

3) τ_b: The Kendall's τ is a statistic used to measure the association between two measured quantities. Specifically, it is a measure of rank correlation [76]:

$$\tau_b = \frac{\sum \hat{c}_{ij} c_{ij}}{\sqrt{\sum \hat{c}_{ij}^2 \sum c_{ij}^2}} \tag{13}$$

where $\hat{c}_{ij}$ is +1 if $\mathcal{O}(\hat{y}_i)$ is greater than $\mathcal{O}(\hat{y}_j)$, 0 if $\mathcal{O}(\hat{y}_i)$ and $\mathcal{O}(\hat{y}_j)$ are the same, and -1 if $\mathcal{O}(\hat{y}_i)$ is less than $\mathcal{O}(\hat{y}_j)$, and similar for $\mathcal{O}(y)$. τ_b values are between -1 and 1.

4) MMAE (The maximum MAE value of all the classes): MMAE is the MAE value of the class with higher distance from the true values to the predicted ones [77]:

$$\text{MMAE} = \max\{\text{MAE}_j; j = 1, \ldots, J\} \tag{14}$$

where MAE_j is the MAE value for the jth class. MMAE values are between 0 and $J - 1$.

These measures aim to evaluate different aspects that can be taken into account when an ordinal regression problem is considered: accuracy measures, patterns are generally well classified; the MAE measures, the classifier tends to predict a class as close to the real class as possible; the τ_b measures, the correlation between predicted and real target pairs; and, finally, the MMAE measures, the order in the class which is worst ordered. The τ_b measure is independent of the values chosen for the ranks representing the classes.

D. Results

The NCL method has been compared with the well-known nominal classification, ordinal regression, and regression techniques given in Section V-B, using the ordinal regression metrics defined in Section V-C. Tables IV and V show the results obtained with the different techniques tested using the training and generalization set. It is important to note that rNN and NCL are nondeterministic methods because they are based on randomly generated numbers. For this reason, these methods were run 30 times, and the best individual in the training set was extracted (rNN_{Best} and NCL_{Best}). Tables IV and V include the average, the standard deviation, and the best values of the C, MAE, τ_b, and MMAE on the training (C_T, MAE_T, τ_{b_T} and MMAE_T) and the generalization (C_G, MAE_G, τ_{b_G} and MMAE_G) sets of these 30 models, together with the results of the remaining methods.

A descriptive analysis of the results leads to the following remarks: the NCL method obtained the best result in terms of C, MAE, τ_b, and MMAE both in training and generalization sets comparing all techniques. As can been observed, the POM model is not able to reflect nonlinear relationships among input variables, which are necessary to perform a realistic classification task. The results obtained for the MMAE metric show that the classifiers tested tend to predict a class quite close to the real class for all the classes.

TABLE IV
TRAINING RESULTS OF THE C_T, MAE_T, τ_{b_T}, AND MMAE_T OF THE METHOD PROPOSED COMPARED WITH THOSE OBTAINED USING DIFFERENT STATISTICAL AND ARTIFICIAL INTELLIGENCE METHODS

	Training Results			
	C_T	MAE_T	τ_{b_T}	MMAE_T
MLR	87.654	0.123	0.922	0.600
RBFN	96.296	0.037	0.977	0.100
MLP	*98.765*	*0.012*	*0.992*	*0.083*
SVC	92.592	0.074	0.953	0.300
POM	82.719	0.185	0.898	0.666
SVMRank	96.296	0.037	0.977	0.100
SVOR(EX)	95.061	0.049	0.969	0.100
SVOR(IM)	**100.000**	**0.000**	**1.000**	**0.000**
ASAOR	97.530	0.022	0.987	0.080
MCOSvm	**100.000**	**0.000**	**1.000**	**0.000**
rNN	$97.121_{0.341}$	$0.021_{0.012}$	$0.974_{0.016}$	$0.092_{0.009}$
NCL	$100.000_{0.000}$	$0.000_{0.000}$	$1.000_{0.000}$	$0.000_{0.000}$
rNN_{Best}	**100.000**	**0.000**	**1.000**	**0.000**
NCL_{Best}	**100.000**	**0.000**	**1.000**	**0.000**

The best result is in bold face and the second best result in italics

TABLE V
GENERALIZATION RESULTS OF THE C_G, MAE_G, τ_{b_G}, AND MMAE_G OF THE METHOD PROPOSED COMPARED WITH THOSE OBTAINED USING DIFFERENT STATISTICAL AND ARTIFICIAL INTELLIGENCE METHODS

	Generalization Results			
	C_G	MAE_G	τ_{b_G}	MMAE_G
MLR	70.370	0.333	0.851	0.666
RBFN	74.074	0.296	0.836	*0.444*
MLP	70.370	0.296	0.860	**0.333**
SVC	70.370	0.370	0.821	1.333
POM	62.965	0.407	0.816	0.666
SVMRank	70.370	0.296	0.868	0.666
SVOR(EX)	70.370	0.296	0.866	**0.333**
SVOR(IM)	74.074	0.259	*0.879*	**0.333**
ASAOR	66.667	0.370	0.821	0.666
MCOSvm	77.777	0.259	0.890	0.444
rNN	$73.898_{1.316}$	$0.2561_{0.033}$	$0.863_{0.045}$	$0.402_{0.062}$
NCL	$83.465_{0.919}$	$0.1592_{0.0190}$	$0.912_{0.019}$	$0.431_{0.093}$
rNN_{Best}	*77.777*	*0.222*	0.873	**0.333**
NCL_{Best}	**85.185**	**0.148**	**0.926**	**0.333**

The best result is in bold face and the second best result in italics

Fig. 5 shows the box plot obtained with the results of the different algorithms in C_G, MAE_G, τ_{b_G}, and MMAE_G. Box plots depict algorithm results according to the smallest observation, lower quartile, median, upper quartile, and largest observation. As seen in Table V and Fig. 5, the NCL method obtains the best result in terms of C_G, MAE_G, and τ_{b_G} out of all the techniques compared. The differences in C_G, MAE_G, and τ_{b_G} are really important with respect to techniques such as MLP, SVC, POM, SVOR(EX), or ASAOR. In general, these results show that the proposed approach based on rNN ensembles is robust enough to classify the sovereign ratings reported by the S&P international CRA, obtaining better results than the rNN base method and state-of-the-art classification algorithms.

E. Discussion

In order to justify the proposal, we have evaluated the projections of the best proposed model (NCL) both in training and test sets. As shown in Fig. 6, most of the patterns (85%) have been correctly classified by the NCL model proposed. On analyzing the four errors committed in the classification test, they

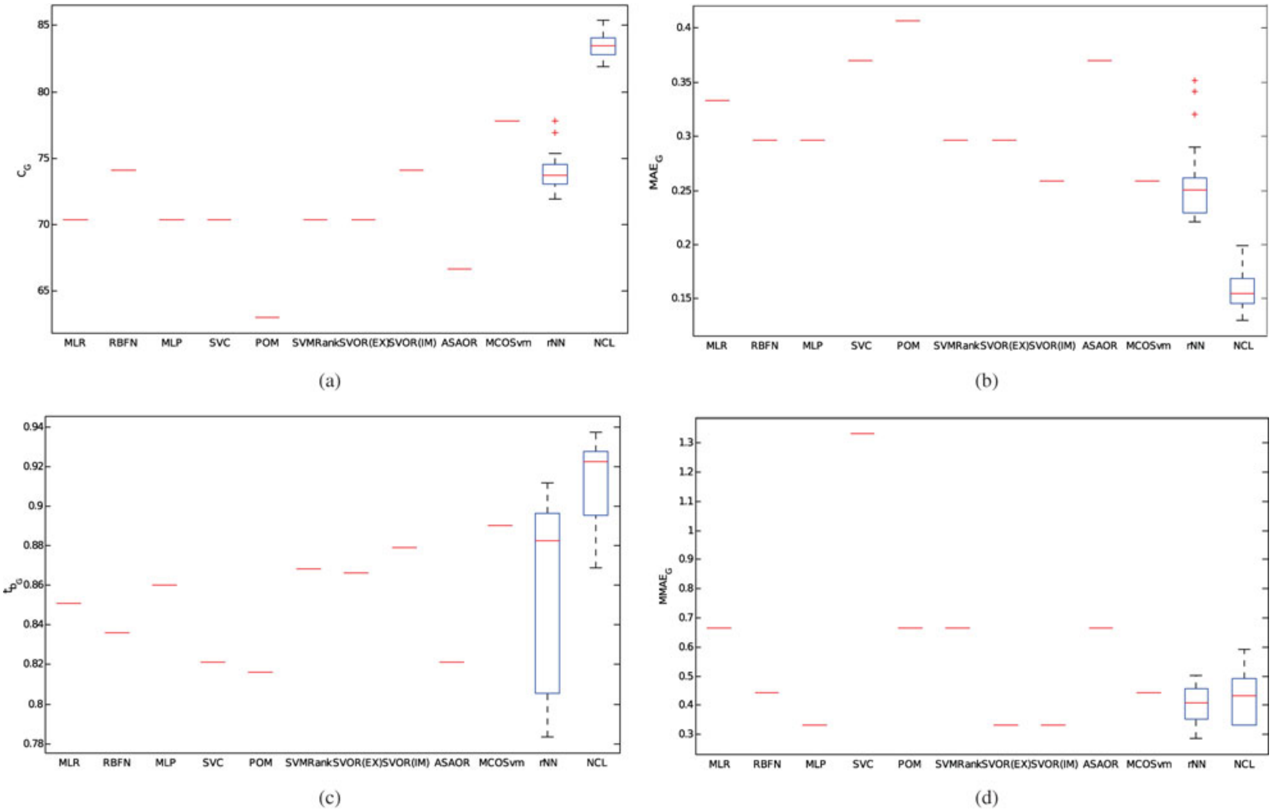


Fig. 5. Box plots: Results of the MLR, RBFN, MLP, SVC, POM, SVMRank, SVOR(EX), SVOR(IM), ASAOR, MCOSvm, rNN, and the NCL method proposed.

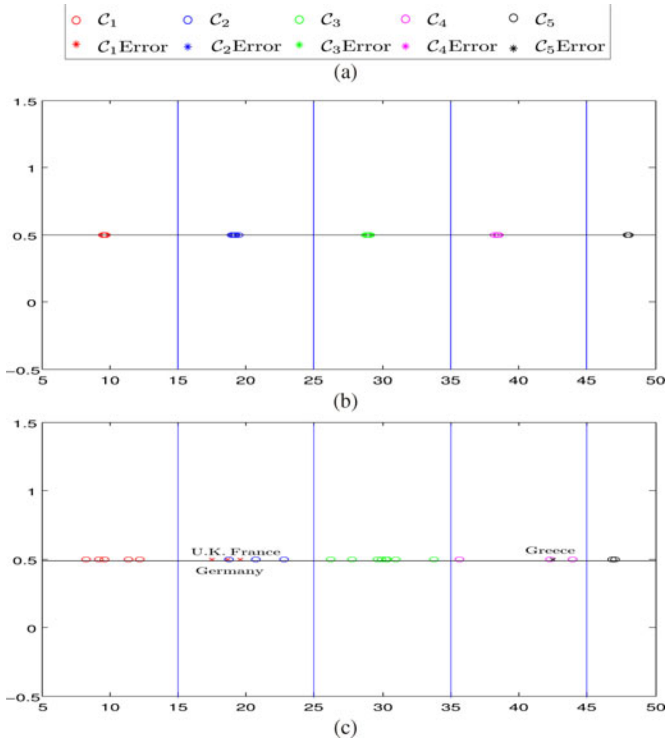


Fig. 6. NCL Projections on country risk ordinal classification datasets. Classes are shown as crosses and dots depending if the pattern has been correctly classified or not. (a) Class Naming. (b) Projection patterns on the real line generated by NCL on the training set. (c) Projection patterns on the real line generated by NCL on the test set.

are seen to correspond to these four countries: Germany, the U.K., France, and Greece. The first three countries have been classified into a lower category than S&P's rating, while Greece has been classified into a higher category than the benchmark.

These differences may be due to the fact that quantitative data employed in our model only provide information on the historical performance of the economy and on its fundamental structural features. They are basically backward-looking, while sovereign rating analysis requires forward-looking evaluations of the risk default over a medium to long-term time horizon. Therefore, examination of past experience has to be supplemented by medium-term projections and by the construction of a range of scenarios that test the vulnerability of a country's economic, political, and financial situation to a variety of shocks generated both internally and externally. In addition, qualitative and judgmental aspects of analysis are unavoidable even in the interpretation of quantitative indicators [78]. Furthermore, in the case of Greece, the reliability of the data employed [79] may be the cause of its better performance.

Finally, in order to make sure that the crisis has not a huge impact on the performance of the model, one additional experiment has been carried out. In this experiment, the performance of the best individual in the training set of the NCL method ($\mathrm{NCL}_{\mathrm{Best}}$) has been tested with two datasets. In these datasets, the training set has been built with the latest years and the generalization set with the earliest years. Two possible configuration has been included: 1) training the model with the 2008, 2009, and 2010 years and validate it with the 2007 year; and 2) training

TABLE VI

TRAINING AND GENERALIZATION RESULTS OF THE C, MAE, τ_b, AND MMAE OF THE NCL$_{\mathrm{Best}}$ METHOD CONSIDERING TRAINING SETS WITH THE LATEST YEARS AND GENERALIZATION SETS WITH THE EARLIEST YEARS

	Training Results (NCL$_{\mathrm{Best}}$)			
	C$_T$	MAE$_T$	τ_{b_T}	MMAE$_T$
{2008, 2009, 2010}	100.000	0.000	1.000	0.000
{2009, 2010}	100.000	0.000	1.000	0.000
	Generalization Results (NCL$_{\mathrm{Best}}$)			
	C$_G$	MAE$_G$	τ_{b_G}	MMAE$_G$
{2007}	81.481	0.222	0.918	0.500
{2007, 2008}	79.630	0.240	0.893	0.833

the model with the 2009 and 2010 years and validate it with the 2007 and 2008 years.

Table VI includes the results of the experiment. As can be observed, the performance of the model in the first case has been slightly reduced (considering the generalization set). In the second case, the degradation of the model performance is higher. The reason for such underperformance is not so much due to the crisis in itself, but to the fact that the second model has been trained using a set which does not include the data of 2008 (included in the generalization set). As shown in Table III, the year 2008 registered a higher number of rating movements than other years, and hence, it provided valuable information for training process.

VI. CONCLUSION

In this paper, we have presented a new ensemble model for ordinal regression problems based on NCL. The models are trained taking into account the correlation in the projections of different individuals in the ensemble. As a result, many accurate and diverse projections are obtained. This paper has tested this approach in a sovereign credit-rating dataset, formed by macroeconomic fundamentals of the 27 EU countries in the periods 2007–2009 (training data) and 2010 (test data), and the corresponding rating variable provided by S&P. Furthermore, an additional experiment has been carried out in order to assess the possible impact of crisis on the best model performance.

The results obtained show that the approaches proposed, which are based on ensembles of MLPs trained with NCL, are robust enough to tackle the ordinal classification of sovereign credit rating provided by the S&P rating agency, and obtain better results than the majority of existing alternative methods. Therefore, it appears that NCL could be a useful tool for informing and supporting the analyst in this decision process, entailing less time and cost. Despite this, such models cannot replace the analyst's role and the partly subjective process of assigning a credit rating.

In future work, this study can be extended by comparing other new methods of ensemble learning for ordinal regression, and by adding the ratings assigned by the other two well-established CRAs, namely, Moody's and Fitch, in order to get a broader view of sovereign credit-rating assessment, since this information is especially valuable in a context of uncertainty in global financial and economic crises and keeping in mind the increasing relevance these agencies are acquiring in influencing the financial market.

REFERENCES

[1] R. Cantor and F. Packer, "Differences of opinion and selection bias in the credit rating industry," *J. Bank. Finance*, vol. 21, no. 10, pp. 1395–1417, 1997.

[2] Y. Jafry and T. Schuermann, "Measurement, estimation and comparison of credit migration matrices," *J. Bank. Finance*, vol. 28, no. 11, pp. 2603–2639, 2004.

[3] P. Behr and A. Güttler, "The informational content of unsolicited ratings," *J. Bank. Finance*, vol. 32, no. 4, pp. 587–599, 2008.

[4] R. Arezki, B. Candelon, and A. Sy, "Sovereign rating news and financial markets spillovers: Evidence from the European debt crisis," IMF Working Paper, 11/69, Mar. 2011.

[5] E. I. Altman and H. A. Rijken, "How rating agencies achieve rating stability," *J. Bank. Finance*, vol. 28, no. 11, pp. 2679–2714, 2004.

[6] J. D. Amato and C. H. Furfine, "Are credit ratings procyclical?," *J. Bank. Finance*, vol. 28, no. 11, pp. 2641–2677, 2004.

[7] A. Duff and S. Einig, "Understanding credit ratings quality: Evidence from UK debt market participants," *Brit. Account. Rev.*, vol. 41, pp. 107–119, 2009.

[8] R. Al-Sakka and O. ap Gwilym, "Split sovereign ratings and rating migrations in emerging economies," *Emerg. Markets Rev.*, vol. 11, no. 2, pp. 79–97, 2010.

[9] M. A. Ferreira and P. M. Gama, "Does sovereign debt ratings news spill over to international stock markets?," *J. Bank. Finance*, vol. 31, no. 10, pp. 3162–3182, 2007.

[10] S. Kim and E. Wu, "Sovereign credit ratings, capital flows and financial sector development in emerging markets," *Emerg. Markets Rev.*, vol. 9, no. 1, pp. 17–39, 2008.

[11] R. Brooks, R. W. Faff, D. Hillier, and J. Hillier, "The national market impact of sovereign rating changes," *J. Bank. Finance*, vol. 28, no. 1, pp. 233–250, 2004.

[12] V. Hooper, T. Hume, and S.-J. Kim, "Sovereign rating changes—Do they provide new information for stock markets?," *Econom. Syst.*, vol. 32, no. 2, pp. 142–166, 2008.

[13] R. M. Cantor and F. Packer, "Determinants and impact of sovereign credit ratings," *Econom. Policy Rev.*, pp. 37–53, Oct. 1996.

[14] E. Durbin and D. Ng, "The sovereign ceiling and emerging market corporate bond spreads," *J. Int. Money Finance*, vol. 24, no. 4, pp. 631–649, 2005.

[15] J. R. M. Hand, R. W. Holthausen, and R. W. Leftwich, "The effect of bond rating agency announcements on bond and stock prices," *J. Finance*, vol. 47, no. 2, pp. 733–752, 1992.

[16] A. Afonso and R. M. Sousa, "The macroeconomic effects of fiscal policy," *Appl. Econom.*, vol. 44, no. 34, pp. 4439–4454, 2012.

[17] I. D. Dichev and J. Piotroski, "The long-run stock returns following bond ratings changes," *J. Finance*, vol. 56, no. 1, pp. 173–203, 2001.

[18] A. Gande and D. C. Parsley, "News spillovers in the sovereign debt market," *J. Financ. Econom.*, vol. 75, no. 3, pp. 691–734, 2005.

[19] R. Alsakka and O. ap Gwilym, "Rating agencies' signals during the European sovereign debt crisis: Market impact and spillovers," *J. Econom. Behavior Org.*, vol. 85, pp. 144–162, Jan. 2013.

[20] A. Afonso, "Understanding the determinants of sovereign debt ratings: Evidence for the two leading agencies," *J. Econom. Finance*, vol. 27, pp. 56–74, 2003.

[21] A. W. Butler and L. Fauver, "Institutional environment and sovereign credit ratings," *Financ. Manage.*, vol. 35, no. 3, pp. 53–79, 2006.

[22] N. Sargen, "Economic indicators and country risk appraisal," *Econom. Rev.*, vol. 1, pp. 19–35, 1977.

[23] Y.-T. Hu, R. Kiesel, and W. Perraudin, "The estimation of transition matrices for sovereign credit ratings," *J. Bank. Finance*, vol. 26, no. 7, pp. 1383–1406, 2002.

[24] E. Bissoondoyal-Bheenick, "Rating timing differences between the two leading agencies: Standard and Poor's and Moody's," *Emerg. Markets Rev.*, vol. 5, no. 3, pp. 361–378, 2004.

[25] A. Afonso, P. Gomes, and P. Rother, "Ordered response models for sovereign debt ratings," *Appl. Econom. Lett.*, vol. 16, no. 8, pp. 769–773, 2009.

[26] J. A. Bennell, D. Crabbe, S. Thomas, and O. ap Gwilym, "Modelling sovereign credit ratings: Neural networks versus ordered probit," *Exper. Syst. Appl.*, vol. 30, no. 3, pp. 415–425, 2006.

[27] J. Yim and H. Mitchell, "Comparison of country risk models: Hybrid neural networks, logit models, discriminant analysis and cluster techniques," *Exper. Syst. Appl.*, vol. 28, no. 1, pp. 137–148, 2005.

[28] K.-J. Kim and H. Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach," *Comput. Oper. Res.*, vol. 39, no. 8, pp. 1800–1811, 2012.

[29] C. Hung and J.-H. Chen, "A selective ensemble based on expected probabilities for bankruptcy prediction," *Exper. Syst. Appl.*, vol. 36, no. 3, part 1, pp. 5297–5303, 2009.

[30] L. Yu, S. Wang, and K. K. Lai, "Credit risk assessment with a multistage neural network ensemble learning approach," *Exper. Syst. Appl.*, vol. 34, no. 2, pp. 1434–1444, 2008.

[31] P. McCullagh, "Regression models for ordinal data," *J. R. Stat. Soc. Series B*, vol. 4, pp. 109–142, 1980.

[32] A. Agresti, *Analysis of Ordinal Categorical Dat*, (Wiley Series in Probability and Statistics). New York, NY, USA: Wiley, 1984.

[33] K. Crammer and Y. Singer, "Online ranking by projecting," *Neural Comput.*, vol. 17, pp. 145–175, 2005.

[34] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems 15*. S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA, USA: MIT Press, 2002, pp. 937–944.

[35] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1019–1041, 2005.

[36] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, pp. 792–815, 2007.

[37] J. S. Cardoso and J. F. P. da Costa, "Learning to classify ordinal data: The data replication method," *J. Mach. Learn. Res.*, vol. 8, pp. 1393–1429, 2007.

[38] B.-Y. Sun, X.-M. Zhang, and W.-B. Li, "An improved ordinal regression approach with sum-of-margin principle," in *Proc. 6th Int. Conf. Natural Comput.*, Aug. 10–12, 2010, pp. 853–857.

[39] E. K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Mach. Learn.*, vol. 65, no. 1, pp. 247–271, 2006.

[40] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging, boosting, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2011.

[41] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proc. Eur. Conf. Mach. Learn.*, 2001, pp. 145–156.

[42] W. Waegeman and L. Boullart, "An ensemble of weighted support vector machines for ordinal regression," *Int. J. Comput. Syst. Sci. Eng.*, vol. 3, no. 1, pp. 47–51, 2009.

[43] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Netw.*, vol. 12, pp. 1399–1404, 1999.

[44] Y. Liu, X. Yao, and T. Higuchi, "Ensembles with negative correlation learning," *IEEE Trans. Evol. Comput.*, vol. 4, no. 4, pp. 380–387, Nov. 2000.

[45] G. Kaminsky and S. L. Schmukler, "Emerging market instability: Do sovereign ratings affect country risk and stock returns?," *World Bank Econom. Rev.*, vol. 16, no. 2, pp. 171–195, 2002.

[46] C. J. Frank and W. R. Cline, "Measurement of debt servicing capacity: An application of discriminant analysis," *J. Int. Econom.*, vol. 1, no. 3, pp. 327–344, 1971.

[47] W. T. Carleton and E. M. Lerner, "Statistical credit scoring of municipal bonds," *J. Money, Credit Bank.*, vol. 1, no. 4, pp. 750–764, Nov. 1969.

[48] F. Fernández-Navarro, C. Hervás-Martínez, P. A. Gutierrez, and M. Carboreno, "Evolutionary q-Gaussian radial basis functions neural networks for multi-classification," *Neural Netw.*, vol. 24, no. 7, pp. 779–784, 2011.

[49] A. Castaño, F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutierrez, "Neuro-logistic models based on evolutionary generalized radial basis function for the microarray gene expression classification problem," *Neural Process. Lett.*, vol. 34, no. 2, pp. 117–131, 2011.

[50] F. Fernández-Navarro, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutierrez, "MELM-GRBF: A modified version of the extreme learning machine for generalized radial basis function neural networks," *Neurocomputing*, vol. 74, no. 16, pp. 2502–2510, 2011.

[51] J. C. Singleton and A. J. Surkan, "Bond rating with neural networks," presented at the Proc. Neural Netw. Capital Markets, London, U.K., 1993.

[52] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study," *Decis. Supp. Syst.*, vol. 37, pp. 543–558, 2004.

[53] D. Trigueiros and R. Taffler, "Neural networks and empirical research in accounting," *Account. Bus. Res.*, vol. 26, no. 4, pp. 347–355, 1996.

[54] C. P. Lim and R. F. Harrison, "Online pattern classification with multiple neural network systems: An experimental study," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 33, no. 2, pp. 235–247, May 2003.

[55] W. Lin, Y. H. Hu, and C.-F. Tsai, "Machine learning in financial crisis prediction: A survey," *IEEE Trans. Syst., Man, Cybern. C. Appl. Rev.*, vol. 42, no. 4, pp. 421–436, Jul. 2012.

[56] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognit.*, vol. 29, pp. 341–348, 1996.

[57] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connect. Sci.*, vol. 8, no. 3, pp. 385–404, 1996.

[58] H. Chen, P. Tino, and X. Yao, "Probabilistic classification vector machines," *IEEE Trans. Neural Netw.*, vol. 20, no. 6, pp. 901–914, Jun. 2009.

[59] G. Brown, J. L. Wyatt, and P. Tiňo, "Managing diversity in regression ensembles," *J. Mach. Learn. Res.*, vol. 6, pp. 1621–1650, 2005.

[60] S. Wang, H. Chen, and X. Yao, "Negative correlation learning for classification ensembles," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2010, pp. 1–8.

[61] M. E. Blume, F. Lim, and A. C. Mackinlay, "The declining credit quality of U.S. corporate debt: Myth or reality," *J. Financ. Econom.*, vol. 53, pp. 458–472, 1998.

[62] E. Gonis and P. Taylor, "Changing credit rating standards in the UK: Empirical evidence from 1999 to 2004," *Appl. Financ. Econom.*, vol. 19, no. 3, pp. 213–225, 2009.

[63] G. Ferri, L.-G. Liu, and J. E. Stiglitz, "The procyclical role of rating agencies: Evidence from the East Asian crisis," *Econom. Notes*, vol. 28, no. 3, pp. 335–355, 1999.

[64] M. Gartner, B. Griesbach, and F. Jung, "Pigs or lambs? The European sovereign debt crisis and the role of rating agencies," *Int. Adv. Econom. Res.*, vol. 17, pp. 288–299, 2011.

[65] D. Beers and M. Cavanaugh, "Sovereign credit ratings: A primer," *Standard Poor's*, 2005.

[66] D. Ratha, P. K. De, and S. Mohapatra, "Shadow sovereign ratings for unrated developing countries," *World Dev.*, vol. 39, no. 3, pp. 295–307, 2011.

[67] G. Feder and L. V. Uy, "The determinants of international creditworthiness and their policy implications," *J. Policy Model.*, vol. 7, no. 1, pp. 133–156, 1985.

[68] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learn.*, vol. 59, no. 1–2, pp. 161–205, 2005.

[69] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Technique*, (ser. Data Management Systems), 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2005.

[70] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1999.

[71] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[72] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst. 19*, 2007, pp. 865–872.

[73] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 145–152.

[74] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, pp. 792–815, 2007.

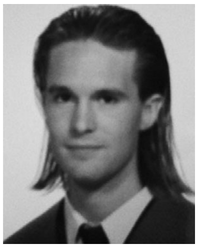
[75] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[76] M. Kendall, *Rank Correlation Methods*. London, U.K.: Griffin, 1948.

[77] M. Cruz-Ramirez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutiérrez, "A preliminary study of ordinal metrics to guide a multi-objective evolutionary algorithm," presented at the 11th Int. Conf. Intell. Syst. Des. Appl., Cordoba, Spain, 2011.

[78] E. Bissoondoyal-Bheenick, "An analysis of the determinants of sovereign ratings," *Global Finance J.*, vol. 15, no. 3, pp. 251–280, 2005.

[79] B. Rauch, M. Gottsche, G. Brahler, and S. Engel, "Fact and fiction in EU-governmental economic data," *German Econom. Rev.*, vol. 12, no. 3, pp. 243–255, 2011.



Francisco Fernández-Navarro was born in Córdoba, Spain, in 1985. He received the B.S. degree in computer science from the University of Córdoba, Córdoba, in 2008, and the Ph.D. degree in computer science and artificial intelligence from the University of Málaga, Málaga, Spain, in 2011.

He is currently a Research Fellow in Computational Management with the European Space Research and Technology Centre (ESTEC), European Space Agency (ESA), Noordwijk, The Netherlands. His current interests include radial basis functions neural networks, evolutionary computation, and hybrid algorithms.

Pilar Campoy-Muñoz was born in Almeria, Spain, in 1982. He received the B.A. degree in business administration and the M.A. degree in research methods on economics and business sciences, both from ETEA, Business Administration Faculty, University of Córdoba, Córdoba, Spain, in 2005 and 2011, respectively, where she is currently working toward the Ph.D. degree with the Department of Economics, ETEA.

She is also a Research Assistant with the Department of Economics, ETEA. Her current research interests include computational economics and financial economics, especially regarding to European Union countries.

Mónica-de la Paz-Marín was born in Córdoba, Spain, in 1973. She received the B.S. degree in business administration from ETEA, Faculty of Economic and Business Sciences, University of Córdoba, Córdoba, in 1996. She also has the equivalent to a Bachelor of Law Degree (B.L.) from the Faculty of Law, University of Córdoba. She is currently working toward the Ph.D. degree in the European Union Doctoral Programme (ETEA-LOYOLA, Department of Quantitative Methods) as well as the M.S. degree in research methods on economics and business sciences with the ETEA-LOYOLA, Faculty of Economic and Business Sciences, Loyola Andalusia University, Seville, Spain.

She initiated and continues her professional activities with the University of Córdoba and as an Adviser in the Regional Government of Andalusia (Regional Ministry of Economy, Innovation and Science, Provincial Delegation). Her research interests are focused on the assessment of research, technological, development, and innovation performance of European Public Policies through a variety of techniques such as data envelopment analysis methodology, artificial neural networks, and modeling Bayesian network causal models as a tool for public policy decision-making support.

César Hervás-Martínez (M'00) was born in Cuenca, Spain. He received the B.S. degree in statistics and operating research from the Universidad Complutense, Madrid, Spain, in 1978, and the Ph.D. degree in mathematics from the University of Seville, Seville, Spain, in 1986.

He is currently a Professor with the Department of Computing and Numerical Analysis, University of Córdoba, Córdoba, Spain, in the area of computer science and artificial intelligence and an Associate Professor with the Department of Quantitative Methods, School of Economics. His current research interests include neural networks, evolutionary computation, and the modeling of natural systems.

Xin Yao (M'91–SM'96–F'03) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 1982, the M.Sc. degree from the North China Institute of Computing Technology, Beijing, China, in 1985, and the Ph.D. degree from USTC in 1990, all in computer science.

From 1985 to 1990, he was an Associate Lecturer and Lecturer with USTC, while working towards the Ph.D. degree on simulated annealing and evolutionary algorithms. He was a Postdoctoral Fellow with the Computer Sciences Laboratory, Australian National University, Canberra, A.C.T., Australia, in 1990, and continued his research on simulated annealing and evolutionary algorithms. He joined the Knowledge-Based Systems Group, Commonwealth Scientific and Industrial Research Organisation, Division of Building, Construction and Engineering, Melbourne, Vic., Australia, in 1991, working primarily on an industrial project on automatic inspection of sewage pipes. He returned to Canberra in 1992 to take up a lectureship with the School of Computer Science, University College, University of New South Wales, Australian Defence Force Academy, where he was later promoted to a Senior Lecturer and Associate Professor. He joined the University of Birmingham, Birmingham, U.K., as a Professor (Chair) of Computer Science in 1999. He is currently the Director of the Centre of Excellence for Research in Computational Intelligence and Applications, University of Birmingham, and a Changjiang (Visiting) Chair Professor (Cheung Kong Scholar) with the USTC. His major research interests include evolutionary artificial neural networks, automatic modularization of machine learning systems, evolutionary optimization, constraint handling techniques, computational time complexity of evolutionary algorithms, coevolution, iterated prisoner's dilemma, data mining, and real-world applications. He has more than 300 refereed publications.

Dr. Yao was the Editor-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION (2003–2008), an Associate Editor or Editorial Board Member of 12 other journals, and the Editor of the *World Scientific Book Series on Advances in Natural Computation*. He has given more than 50 invited keynote and plenary speeches at conferences and workshops worldwide. He received the President's Award for Outstanding Thesis by the Chinese Academy of Sciences for his doctoral work on simulated annealing and evolutionary algorithms in 1989. He received the 2001 IEEE Donald G. Fink Prize Paper Award for his work on evolutionary artificial neural networks.