Logistic evolutionary product-unit neural network classifier: the case of agrarian efficiency

Article *in* Progress in Artificial Intelligence · December 2015

DOI: 10.1007/s13748-015-0068-7

5 authors, including:

Mercedes Torres Jiménez
Universidad Loyola Andalucía
24 PUBLICATIONS **308** CITATIONS
SEE PROFILE

Javier Sánchez-Monedero
University of Cordoba (Spain)
65 PUBLICATIONS **1,229** CITATIONS
SEE PROFILE

Salud Millán Lara
Universidad Loyola Andalucía
17 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Cesar Hervás Martínez
University of Cordoba (Spain)
338 PUBLICATIONS **6,824** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Logistic evolutionary product unit neural network classifier: The case of agrarian efficiency

Mercedes Torres-Jiménez · Carlos R. García-Alonso · Javier Sánchez-Monedero · Salud Millán · César Hervás-Martínez

Received: date / Accepted: date

Abstract By using a high-variability sample of real agrarian enterprises previously classified into two classes (efficient and inefficient), a comparative study was carried out to demonstrate the classification accuracy of logistic regression algorithms based on evolutionary productunit neural networks. Data envelopment analysis considering variable returns-to-scale (BBC-DEA) was chosen to classify selected farms (220 olive tree farms in dry farming) as efficient or inefficient by using surveyed socio-economic variables (agrarian year 2000). Once the sample was grouped by BCC-DEA, easy-to-collect descriptive variables (concerning the farm and farmer) were then used as independent variables in order to find a quick and reliable alternative for classifying agrarian enterprises as efficient or inefficient according to their technical efficiency. Results showed that our proposal is very promising for the classification of complex structures (farms).

Keywords Neural Networks · Logistic Regression · Classification · Product-Unit · Evolutionary Algorithms · Agrarian Technical Efficiency · Data Envelopment Analysis.

This work has been partially subsidized by the TIN2014-54583-C2-1-R project of the Spanish Ministerial Commission of Science and Technology (MINECO), FEDER funds and the P2011-TIC-7508 project of the "Junta de Andalucía" (Spain).

Mercedes Torres-Jiménez, Carlos R. García-Alonso, Javier Sánchez-Monedero and Salud Millán
Loyola-Andalusia University, Department of Mathematics and Engineering
E-mail: {mtorres, cgarcia, jsanchez, smillan}@uloyola.es

César Hervás-Martínez
University of Córdoba, Department of Computer Science and Numerical Analysis. E-mail: chervas@uco.es

1 Introduction

The logistic regression (LR), is a special case of generalized linear model methodology where the assumptions of normality and the constant variance of the residuals are not satisfied. In this paper LR is improved (to include the nonlinear effects of the covariates) taking into account the combination of linear and product-unit models [6,10,20,23]. Product-unit functions (PU) are nonlinear basis functions (mathematical transformations of the input variables) designed using the product of the covariates raised to arbitrary powers (real values). The nonlinear basis functions of the proposed model correspond to a special class of feed-forward neural networks, namely product-unit neural networks (PUNN). Introduced by Durbin and Rumelhart [6] and developed recently where product-unit neural networks (PUNN) express strong covariate interactions. In this way, the LR model can be structured, on one hand, with all covariate product units: logistic regression by the product-unit model (LRPU) or, on the other, with both PU and initial covariates: logistic regression by the initial and the product-units covariates model (LRIPU).

The objective of this paper was to check the accuracy and interpretability of our hybrid classification algorithms as an alternative for classifying sets of observations in an uncertain environment where a great deal of interaction between our input variables was expected. According to this goal, LRPU and LRIPU algorithms were compared with LR and linear discriminant analysis (LDA), selecting a relevant problem in the agrarian economy framework: the productive efficiency of agrarian enterprises (olive tree farms).

The olive farm has been selected not only for its enormous relevance in the Andalusian economy but also because of its role in the sustainable development and in the reduction of the loss of population in the rural areas. Nowadays the European Union (EU) agrarian sector is more or less subsidised, and this circumstance requires it to be socially committed; that is, its financial support should be socially justified in terms of employment, environmental maintenance, food quality, efficiency, best practices and so on [9,22] . So a subsidised farm should aspire to efficiency in a sustainable environment. According to the productive approach to technical efficiency, the farm is mainly devoted to producing outputs to be sold in the market to obtain a financial profit. A productively efficient farm implies using input that reasonably avoids over-utilization [24].

Data Envelopment Analysis (DEA) was applied to determine the technical efficiency of Decision Making Units (DMUs, farms) without previous assumptions like, for example, the knowledge of production functions which is often unknown [1–3,5]. In order to apply DEA it is necessary to know the exact values of all the inputs consumed (i.e. fertilizers, pesticides, etc.) for every DMU as well as the outputs produced (i.e. revenues). However, a key advantage of DEA over other approaches like the econometric Stochastic Production Frontier (SPF) is that DEA does not require any pre-described structural relationship between the inputs and resultant outputs, so allowing greater flexibility in the frontier estimation. It can also accommodate multiple outputs into the analysis. A disadvantage of the technique, nonetheless, is that it does not account for random variation in the output, and so attributes any apparent shortfall in output to technical inefficiency. In spite of the main objective of this work is focused in forecasting with simple and only a few information the efficiency of a DMU and is not so important the methodology used to compute the technical efficiency.

In this paper, 220 olive-tree farms in dry farming were selected and grouped into efficient and inefficient groups by DEA using the socio-economic variables surveyed (which supposes a great effort in terms of time and economical cost). The division or classification obtained was then used to check our classification algorithms but now considering only easy-to-collect variables describing the structure of the farm and the farmer. Our hypothesis was that it was possible to classify farms according to their efficiency by using only these descriptive and easy-to collect variables as the independent ones(instead of difficult-to-obtain and expensive socio-economic ones). Obviously, all of them were different from those employed in DEA.

The classification results obtained using these variables could prove to be especially relevant for decision makers (politicians, rural development program managers, etc.). They who could use this classification for rural planning purposes: for example, to identify those farms that should be the principle aid recipients due to their superior or inferior productive efficiency and compare them to the rest in order to improve the management and degree of competitiveness of the less efficient ones.

This paper is structured as follows: in Section 2 the classification methods to be used for determining farm efficiency were described; Section 3 briefly describes the DEA model; Section 4 presents the set of variables selection both for DEA model and for the machine learning models, those variables referring to a very relevant and strategic agrarian group in Andalusia (olive-tree farms in dry farming) and summarizes efficiency results in DEA models; classification results are statistically described in Section 5 together with the most relevant findings obtained using our hybrid classification models. Finally, some illustrative conclusions are drawn in Section 6.

2 Classification methods

2.1 Logistic regression with product-unit covariates

In classification of real problems, it is not possible to assume that the generic function for determining the best classifier is always linear. According to that, several approaches for modelling non-linear systems have been proposed recently: the method of fractional polynomials and the method of fitting a generalized additive model.

In this study, based on the latest researched models, we propose a new alternative for a non-linear function $f(\mathbf{x}, \theta)$ by the inclusion of product-unit functions in the structure of the function $f(\mathbf{x}, \theta)$ establishing therein two parts: the first one is linear (the covariables of the LR) and the other, non-linear, made up of covariates formed as product-unit functions, which build the non-linear part of the function defined as:

$$B_j(\mathbf{x}, \mathbf{w}_j) = \left(\prod_{i=1}^{k} x_i^{w_{jl}} \right), \tag{1}$$

where j ranges from 1 to m being m the number nodes in the hidden layer, so that the activation of the j-th node in the hidden layer is given by (1). In this way, a logistic regression by product-units and initial covariates model (LRIPU) is given in matrix form by:

$$f(\mathbf{x}, \theta) = \mathbf{x}^T \alpha + \mathbf{B}^T(\mathbf{x}, \mathbf{W}) \beta \tag{2}$$

where $\mathbf{x} = (1, x_1, \ldots, x_p)^T$ (p being the number of inputs), $\mathbf{B}(\mathbf{x}, \mathbf{W}) = [B_1(\mathbf{x}, \mathbf{w}_1), \ldots, B_m(\mathbf{x}, \mathbf{w}_m)]^T$ with $B_j(\mathbf{x}, \mathbf{w}_j)$ defined in (1), and the parameters $\theta = (\alpha, \beta, \mathbf{W})$, where $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$, $\beta = (\beta_1, \ldots, \beta_m)^T$ and $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_j)$ being $\mathbf{w}_l = (w_{j1}, \ldots, w_{jp})^T$ in which $w_{jl} \in \mathbb{R}$. The LRPU model only includes the second term of (2) (product-units). So the conditional distribution is now:

$$p(\mathbf{x}, \theta) = \frac{\exp(\mathbf{x}^T\alpha + \mathbf{B}^T(\mathbf{x}, \mathbf{W})\beta)}{1 + \exp(\mathbf{x}^T\alpha + \mathbf{B}^T(\mathbf{x}, \mathbf{W})\beta)} \tag{3}$$

In this case, the decision boundaries are generalized surface response models.

We define a pattern as a p-dimensional feature vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$ and a class label $y \in \mathcal{Y}$. A training dataset $\mathbf{D}$ consist on n patterns

$$\mathbf{D} = \{(\mathbf{X}, \mathrm{Y}) = (\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y}\,(i = 1, \ldots, n)\},$$

with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and $y_i \in \{0, 1\}$. We use the method of maximum likelihood to estimate the parameters α and β –second step– because $\mathbf{W}$ was previously estimated by the evolutionary algorithm (EA) in the first step in the linear predictor in the linear predictor $\mathbf{x}^T\alpha + \mathbf{B}^T(\mathbf{x}, \mathbf{W})\beta$). Each sample observation follows a Bernouilli distribution, there, since the observations are independent, the likelihood function is:

$$L(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}, \tag{4}$$

and the negative log-likelihood for those observations is:

$$\ln L(y_1, y_2, \ldots, y_n, \alpha, \theta) = \\ = \sum_{i=1}^{n} \left[y_i f(\mathbf{x}_i, \alpha, \theta) - \ln(1 + e^{f(\mathbf{x}_i, \alpha, \theta)} \right] \tag{5}$$

2.2 The estimation of coefficients

The methodology proposed to estimate both LRPU and LRIPU parameters is a three step procedure based on the combination of global exploration algorithm (an evolutionary algorithm, EA) and a local optimization procedure (carried out by a maximum-likelihood optimization method). First the parameters of the PU are determined by the EA, second, the product-units (nonlinear terms) are added to the lineal LR model and the LR model is trained, and third the covariables of the model are pruned.

The first step consist on the application of the EA to design the structure and learning of the weights of a PU neural network. It begins the search with an initial population, and at each iteration, the population is updated using a population-update algorithm.

The evolutionary process determines the number m of potential basis functions of the model and the corresponding vectors w_j of exponents. The algorithm of our proposed EA is the following (details can be found in [10,11]):

1. Generate a random initial population of size P.
2. Repeat the following steps until the stopping criterion is fulfilled:
   (a) Calculate the fitness of every individual in the population and rank the individuals regarding their fitness.
   (b) The best individual is copied into the new population (elitism).
   (c) The best 10% percent of individuals of the population are replicated and substitute the worst 10% individuals.
   (d) Apply parametric mutation to the best 10% of individuals.
   (e) Apply structural mutation to the remaining 90% of individuals.

For this binary classification problem we consider the mean squared error (MSE) of an individual g of the population as:

$$MSE(g) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{6}$$

being y_i and $\hat{y}_i$ the actual and predicted values (0 or 1 in the binary case), and N the number of training patterns. The output values are estimated after a softmax transformation of the $f(\mathbf{x}, \theta)$ output and the fitness function is defined with a strictly decreasing transformation of the MSE:

$$A(g) = \frac{1}{1 + MSE(g)}. \tag{7}$$

In the present work, the algorithm parameters of the exponents w_{ji} are randomly initialized within the interval $(-1, 1)$ and the coefficients β_{kj} in the interval $(-5, 5)$, the maximum number of hidden nodes is $m = 4$ and the size of the population $P = 1000$. The number of nodes that can be added or removed in a structural mutation is one or two. The number of connections that can be added or removed in a structural mutation is a number from one to six. The stopping criterion is reached whenever one of the following two conditions is fulfilled: i) where there is no improvement during 20 generations either in the average performance of the best 20% of the population or in the fitness of the best individual; or ii) the algorithm achieves 100 generations.

Once the basis functions are determined by the EA, in a second step, we consider a transformation of the input space by adding the nonlinear transformations of the input variables given by the basis functions obtained by the EA. The model is linear in these new variables together with the initial covariates. Numerical search methods could be used to compute maximum likelihood estimates (MLE) $\hat{\alpha}$ and $\hat{\beta}$.

However, it turns out that we can use iteratively reweighed least squares (IRLS) to actually find the MLE. We use the SPSS computer program that implements IRLS for the LR model. In order to define the LR using only product units as covariates, the LRPU model simplifies equation (2) establishing $\alpha = (\alpha_0, 0, \ldots, 0)^T$, in this form, we obtain logistic regression models where the linear and non-linear structure of the function $f(\mathbf{x}, \theta)$ has been modeled only with associated covariates to underlying interactions within the initial covariates.

The third step is a backward-step procedure where the covariates of the model obtained in the second step are pruned sequentially (starting with the full model with all the covariates) until further pruning do not improve the fit. At each phase, the least significant covariate is deleted (the greatest critical $p - value$ in the hypothesis test) to predict the response variable, where the associated coefficient equal to zero is the hypothesis to be contrasted. The procedure finishes when all tests provide $p - values$ smaller than the fixed significance level, α, and the model obtained fits well.

3 Technical efficiency: the DEA model

In order to evaluate the relative performance of a set of decision making units (DMU) that produce multiple outputs consuming multiple inputs, DEA methods are well-known as non-parametric, data-oriented approaches, that have developed greatly since the seminal paper of Charnes et al. [3]. DEA does not need the a priori assumptions associated with other approaches for performance appraisal such as, for example, statistical regression ones. Recent relative efficiency definitions assume that a DMU can be considered 100% efficient in a set of selected DMU if, and only if, according to existing information (inputs and outputs), there is not any real evidence that some inputs or outputs could be improved without worsening any of their inputs or outputs. Based on this definition, many different DEA models have been developed, including the economic concept of returns to scale. Returns to scale can be considered variable when a proportional increase or decrease in all the inputs implies more (increasing returns to scale) or less (decreasing returns to scale) than proportional input increase or decrease. DEA models immediately assume this realistic approach [2,5].

If analyse a set of DMU where each DMU_j, which$(j = 1, 2, ..., n)$ produces identical outputs in different quantities, $y_{rj}(r = 1, 2, ..., s)$ and consumes also identical inputs in different amounts, $x_{ij}(i = 1, 2, ..., m)$, according to the standard variable returns to scale model (BCC-DEA), the technical efficiency of a selected DMU can be evaluated using the primal "envelopment form" using the following linear model:

$$\min \theta_0 - \epsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \tag{8}$$

subject to

$$\theta_0 x_{i0} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- \text{ for } i = 1, \ldots, m \tag{9}$$

$$y_{r0} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \text{ for } r = 1, \ldots, s \tag{10}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \text{ where } \lambda_j, s_i^-, s_r^+ \geq 0, \forall i, r, j, \tag{11}$$

ϵ being a non-Archimedean element smaller than any positive real number, λ_j the model variables and s_i^- and s_r^+ the corresponding slacks.

According to model [9], a DMU is efficient if and only if $\hat{\theta}_0 = 1$ and all slacks are zero. The input-oriented BCC model, that is the most realistic approach in the agrarian sector, analyses the possibility of reducing input consumption to produce the same amount of outputs in every DMU analysed [2,5]. Once the DEA model is solved, it assigns a score to the patterns. Value one means that the farm is technically efficient, and positive values less than one to the non-efficient farms. Then, with DEA scoring we can classify the DMU set (agrarian business) in two groups: efficient ($Y = 1$ or positive class in the classification problem) and inefficient ($Y = 0$ or negative class in the classification problem). In this paper, the relative technical efficiency of olive tree farms in dry farming was calculated using survey-based socio-economic variables (see Table 1).

4 Experimental design: selection and justification of the variable and decisional framework

The samples of real agrarian enterprises were randomly selected to be representative at a provincial level according to the Andalusian distribution of farm sizes and agrarian activities: crops and cattle [9,19,22]. The

Table 1 Socio-economic variables of surveyed olive tree farms to calculate productive efficiency using BCC-DEA (source: CAC 3/2001 project)

#	Variable description	abbrev.	DEA
1	Energy structural costs (electricity and liquid and solid fuels (10^2 €)	SC1	Input
2	Other structural costs except for structural hand labour (10^2 €)	SC2	Input
3	Taxes (10^2 €)	SC3	Input
4	Structural hand labour cost (10^2 €)	SHL	Input
5	Structural revenues including non-agrarian revenues (10^2 €)	SR	Output
6	Family hand labour over total hand labour on the farm (%).	FHL	Input
7	Input costs of the farm crops (10^2 €)	ICO	Input
8	Other costs of the farm crops (10^2 €)	CCO	Input
9	Total hand labour costs of farm crops (10^2 €)	HLC	Input
10	Total crop revenues (10^2 €)	CR	Output
11	Total crop subsidies (10^2 €)	SUB	Input

Table 2 Basic statistics of descriptive variables in total sample

Variable	Mean	Standard Deviation	Variation Coefficient
CA	17.68	27.61	156.00%
PO	87.76	25.61	29.00%
NC	1.40	0.91	65.00%
YL	3.29	2.27	69.00%
HL	45.02	30.51	68.00%
TR	0.75	0.92	123.00%
ST	2.48	1.35	55.00%
AG	2.15	0.89	42.00%
HE	539.9	202.44	37.00%
SL	76.03	25.77	34.00%
ER	37.25	29.81	80.00%

socio-economic structure of these agrarian enterprises, obtained from very detailed survey questionnaires, attempted to achieve the greatest possible precision in determining their productive structure, costs and revenues.

Our classification problem can be structured in three sequential phases: In the first, olive tree farms in the dry farming set from the original databases were selected in order to analyse the potential to reproduce efficiency results obtained from BCC-DEA of LR, LRPU and LRIPU models compared to LDA. This strategic farms group is nowadays one of the most important ones in Andalusia[1]. The sample analysed was made up of 220 complex farms spread all throughout Andalusia. In the second phase, BCC-DEA was chosen to calculate the productive efficiency, relative, of the sample considered. Only socio-economic variables (11 in total) were taken into account as stated in Table 1 (9 DEA inputs, agrarian resources, $\mathbf{x}_{ij}$, and two DEA outputs, $\mathbf{y}_{rj}$, economic results). The set of variables selected to calculate farm efficiency using BCC-DEA are very difficult to collect in the field and relatively difficult to calculate. In the third phase, the classification was carried out using DEA. From a productive point of view, 61 olive tree farms (27.73%) were considered efficient ($Y = 1$). The resulting pre-classification produced a non-balanced structure, a circumstance that added an additional difficulty to the inner statistical variability of the training samples (see Table 2) that is reproduced in the generalization ones.

Once the original set of farms was divided into two groups, efficient or inefficient, a selected easy-to-collect set of variables in the field that describe these farms and the corresponding farmers (see Table 3 for variables description) was considered as the input group of variables. The productive efficiency of each DMU (farm) was the dependent variable, for classification purposes. In the classification process, ten random different training/generalization samples (and hold-out procedure) were designed using approximately 60% of the farms for training and the remainder 40% for generalization..

In order to evaluate and compare the precision of the proposed classification models, the Correct Classification Rate (CCR), Producer's Accuracy (PA) and User's Accuracy (UA) measures were selected [3]. The first one (CCR) can be defined as the percentage of total correct classified observations with respect to the total number of observations. The PA is the number of farms correctly classified as positive class (efficient class, $Y = 1$) with respect to the total number of farms that belongs to that class (generally known as precision). UA is calculated as the number of farms correctly classified as positive class with respect to the total number of farms that was classified as that class by the algorithm (also known as recall). The best classification method, is that

[1] These farms represent 59% of Spanish agricultural land and 27% of that in the EU. Moreover. Andalusia is the main olive-producing region in Spain yielding more than 70% of the total production. There are whole areas devoted to the olive oil sector, which represents 30% of Andalusian agricultural employment.

Table 3 Descriptive variables of surveyed olive-tree farms in dry farming (sources: ^{1}CAC 3/2001 project and ^{2}SIMA [4]

#	Variable description	abbrev.
1	Region. Sample design.	RE^1
2	Total cultivated area of the farm (hectares).	CA^1
3	Percentage of olive tree area over CA (%).	PO^1
4	Number of farm crops.	NC^1
5	Olive tree yield (ton of olives/hectare).	YL^1
6	Percentage of non-family hand labour cost over total production costs of the farm (%).	HL^1
7	Does the manager or the family have non-agrarian revenues? (Yes or No).	NRE^1
8	Number of farm tractors.	TR^1
9	Training level of the farm manager. (1: None, 2: Basic, 3: High school level, 4: Professional training, 5: First university degree, 6: University master o higher degree).	ST^1
10	Does the manager have agrarian studies? (Yes or No).	AS^1
11	Farmer age (<40 years old, between 40 and 55, between 55 and 65 and >65 years old).	AG^1
12	Manager sex (1: Male, 2: Female).	SX^1
13	Does the farm manager sell directly to consumers? (Yes or No).	CON^1
14	Does the farm manager sell directly to wholesalers? (Yes or No).	WH^1
15	Does the farm manager sell directly to retailers? (Yes or No).	RT^1
16	Does the farm manager sell directly to industry? (Yes or No).	IN^1
17	Is the farmer a cooperative member? (Yes or No).	COO^1
18	Average altitude of the farm municipality (meters over sea level).	HE^2
19	Average slope of the farm municipality (%).	SL^2
20	Percentage of agrarian soils in the farm municipality where erosion can be considered moderate (%).	ER^2

where both PA and UA values are equal to one, implying that CCR is equal to one.

5 Results

In order to compare the LDA, LR, LRPU and LRIPU models, values of CCR, PA and UE for training and generalization sets are calculated.

5.1 Analysis of performance of LR, LRPU and LRIPU models

Based on our decisional framework and considering previously and randomly (geographical conglomerates) designed samples, we explain our classification into efficient and non-efficient farms through three logistic models: standard LR, LR with product-unit covariates (LRPU) and LR with product-unit and initial covariates (LRIPU). In this way, we can observe the improvement of the hybrid model LRIPU in the different metrics. The results are summarized in Table 4.

In the majority of the selected ten samples, the CCR was greater than or equal to 70%, which is an excellent classification ratio considering the complexity of the olive-tree farm database (see Table 3).

The best mean results obtained in CCR were reached by LRIPU models 81.2% in training set and 75,2% in the generalization set.

Regarding PA metric, inefficient farms ($Y = 0$) were more easily recognized in general than efficient ones ($Y = 1$). This is a problem that offen rises in imbalanced problems, were the worst performance is achieved for the less populated classes [14, 18]. In the former case (inefficient), the best mean percentage was for LRPU, with values 96.1% and 93.2% (respectively for training and generalization sets). However, for efficient farms, these rates fell to 46.3% and 31.6% with the LRIPU methodology. We can conclude that LRPU was better recognising the inefficient group and LRIPU was better in the efficient group classification (the most difficult patterns to identify). So we can observe that non linear models (LRPU and LRIPU) improve classification of LR for metric PA. Considering results of metric UA (classify inefficient farms), the three methods compared present results greater than 78% in training and a 75.9% in generalization. Considering efficient farms ($Y = 1$), UA results reveals worse performance, over (76-78)% in training and (55-64)% in generalization, which means that the selected algorithms were sensitive to the sample design, probably related to the above-mentioned imbalanced structure of the data analysed. In all the cases the best methodology considering this metric is LRIPU.

Finally, since LRIPU and LRPU have shown the best PA and UA absolute results compared with LR, it makes sense to consider nonlinear variables in the LR model.

Table 4 Mean % of PA, UA and CCR (in 10 hold-out) for training and generalization sets using, LR, LRPU and LRIPU models. (best results in bold)

	Training set PA			Generalization set PA		
Class\Method	LR	LRPU	LRIPU	LR	LRPU	LRIPU
$Y = 0$ Inefficient	95.2	**96.1**	95.2	87.7	**93.2**	91.9
$Y = 1$ Efficient	39.6	33.3	**46.3**	19.2	27.0	**31.6**
	Training set UA			Generalization set PA		
Class\Method	LR	LRPU	LRIPU	LR	LRPU	LRIPU
$Y = 0$ Inefficient	80.2	78.7	**80.5**	75.9	75.9	**77.8**
$Y = 1$ Efficient	76.8	76.4	**77.4**	54.9	63.4	**63.8**
Total Farms	79.7	78.4	**81.2**	68.8	74.4	**75.2**

5.2 Comparison of the proposed method to reference machine learning algorithms

In order to test the reliability of both the LR and product-unit (LRPU and LRIPU) classification methods, a selected set of different methodologies have been chosen to classify our database. All algorithms but LDA (which was applied with SPSS) are part of WEKA machine learning software release 3.4.0. [12]. We have selected a set of popular machine learning methods: LDA [15] a method developed to address the classification problem from a linear multidimensional perspective; IB1, an algorithm belonging the k-Nearest-Neighbour family proposed by [16]; a multilayer Perceptron neural network (MLP) that uses a back-propagation learning algorithm [13]; a neural network technique with normalized Gaussian radial basis functions (RBFNN) [13]; the C4.5 algorithm, which is an extended form of ID3 used for building the decision tree [21]; AdaBoostM1 (ADABM1), an algorithm that uses the boosting procedure to improve the classification accuracy of tree-based classifiers [8]. Logistic Model Trees, (LMT) consist on a standard decision tree structure with logistic regression functions at the leaves [17]; ADTree, which uses the boosting procedure to decision tree algorithms and has been shown to produce very accurate classifiers [8, 7].

Table 5 summarizes the results of several performance metrics for the generalization sets. In all cases, LRPU and LRIPU (our proposed models) showed the best classification results in CCR (in minimum, maximum and mean values). In the inefficient group they were also the best and second best (PAI and UAI values). However, when analysing efficient farms ($Y = 1$), LDA showed better results (in terms of PA) than our approach. LRIPU stands out especially in comparison with LDA, IB1, MLP and RBFN. Considering the standard deviation (SD), the most unstable models in CCR were C4.5, RBFNN and LR.

Statistical tests for mean CCR values of the 10 hold-out samples, were applied in order to assess the statistical significance of the differences observed in the different methods and determine the best classification performance. First of all, a Kolmogorov-Smirnov's test (KS-test) with a significance level $\alpha = 0.05$ was used to evaluate if the different performance metrics in all the methods sample for method follow a normal distribution. Since no method obtained a p-value lower than the critical level for the mean CCR measure a normal distribution cannot be assumed in any of the cases. As a consequence, a non-parametric Friedman's test for dependent samples was selected in order to see if the applied method significantly affected the results obtained. The test concluded that these differences were significant (with a p-value=0.00). So, the statistical analysis ended by applying Wilcoxon's signed-rank test for all pairs of algorithms, results are shown in Table 6. These results include, for each method, the number of times that algorithms statistically outperformed (wins, W), the number of draws (non-significant differences, D) and the number of losses (number of algorithms that outperform the method, L). The conclusion of the statistical test analysis confirm the superiority of LRIPU.

6 Discussion: relevant findings for decision makers

In the present work we have addressed the problem of classifying productive efficiency of olive-three farms in dry farming. We propose to use the DEA method to perform the patterns labelling based on the scored provided by DEA using variables which are difficult or costly to obtain. Then, machine learning models are training using easier to obtain descriptor variables together with the DEA classification. The LRPU and LRIPU models have demonstrated their good accuracy performance in terms of CCR, PA and UA metrics. Both models are among the best when compared to a selected group of different classification algorithms.

Our contribution is relevant from a decision maker (farmers, agrarian policy experts or politicians among others) point of view. We provide a framework to accurately predict efficiency based on a statistical learning

Table 5 Minimum, maximum, mean, standard deviation for CCR; and mean for PA and UA for Inefficient and Efficient farms (best results are marked in bold)

Measure	CCR				PAI	PAE	UAI	UAE
Method	Min.	Max.	Mean	SD.	Mean	Mean	Mean	Mean
ADABM1	67.0	71.3	69.3	1.7	89.3	19.5	73.4	47.8
ADTREE	68.5	68.9	68.8	0.2	78.1	42.2	77.3	42.5
C4.5	65.9	74.3	70.5	4.7	87.9	30.6	77.0	52.8
IB1	61.4	68.5	65.2	2.9	78.1	31.1	74.9	34.9
LDA	59.2	59.8	58.5	0.3	66.8	**43.4**	**79.5**	28.2
LMT	68.0	71.1	69.7	1.3	88.3	25.0	72.6	45.2
LR	64.2	71.1	68.8	3.7	87.7	19.2	74.1	39.2
LRIPU	**71.9**	**76.4**	**75.2**	1.9	**91.9**	31.6	77.9	**59.2**
LRPU	70.9	75.1	74.4	1.5	**93.2**	27.0	76.8	57.3
MLP	62.4	66.3	63.7	1.8	76.8	30.0	74.3	33.0
RBFNN	59.1	72.0	66.9	5.6	87.2	14.2	72.7	27.6

Table 6 Number of wins (W), draws (D) and losses (L) when comparing the different methods using the Wilcoxon's signed rank test with $\alpha = 0.10$

	ADABM1	ADTREE	C4.5	IB1	LDA	LMT	LRIPU	LRPU	MLP	LR	RBFNN
Rank mean	6.00	5.50	9	4.25	1.25	7	12.00	10.00	3	8	4.50
W	2	2	2	1	0	2	10	7	1	1	0
D	6	6	7	7	1	6	0	3	2	8	7
L	2	2	1	2	9	2	0	0	7	1	3

methodology and easily collected variables avoiding the necessity of complex and expensive survey processing needed to obtain socio-economical variables at micro-economy level. More precisely, the determination of productive efficiency is critical in terms of rural sustainability because it concerns a reasonable input (resources) consumption in the rural environment.

Based on statistical evidences, we can confirm that the best models are LRPU and LRIPU, being very precise in classifying productively inefficient farms. Nevertheless, those models have a performance decay in detecting efficient farms. However, in our specific case, the study and detection of inefficient farms is more useful than the detection of efficient farms. The interest in detecting inefficient farms in the context of agrarian sustainability studies arises in the problematic situation from productive-financial and social points of view. Inefficient farms do not manage input consumption appropriately and can contribute to processes of erosion, pollution among others.

In addition to the binary classification, BCC-DEA provides richer information when studying an inefficiency-efficiency ratio less than one $[0, 1)$, that is, degrees of inefficiency can be analysed rather than absolute classifications. On the other hand, BCC-DEA since classifies a specific DMU as efficient by simply appointing a 1, it is very difficult to distinguish efficiency levels within the efficient group of farms (the super-efficiency problem). In this direction, as future work, BCC-DEA information could be used to addressing the problems as a multi-class ordinal classification problem, to have a richer farms groups.

References

1. de Andrés, J., Landajo, M., Lorca, P.: Forecasting business profitability by using classification techniques: A comparative analysis based on a Spanish case. European Journal of Operational Research **167**(2), 518 – 542 (2005)
2. Banker, R.D., Cooper, W.W., Seiford, L.M., Thrall, R.M., Zhu, J.: Returns to scale in different DEA models. European Journal of Operational Research **154**(2), 345 – 362 (2004). {DEA} and its uses in different countries
3. Charnes, A., Cooper, W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research **2**(6), 429 – 444 (1978)
4. Consejería de Economía y Hacienda de la Junta de Andalucía: Sistema de Información Multiterritorial de Andalucía, Instituto de Estadística de Andalucía (2005). URL http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/
5. Cooper, W., Seiford, L., Zhu, J.: Data envelopment analysis. In: W. Cooper, L. Seiford, J. Zhu (eds.) Handbook on Data Envelopment Analysis, *International Series in Operations Research & Management Science*, vol. 71, pp. 1–39. Springer US (2004)
6. Durbin, R., Rumelhart, D.: Products units: A computationally powerful and biologically plausible extension to backpropagation networks. Neural Computation **1**(1), 133–142 (1989)
7. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceedings of the Sixteenth International Conference on Machine Learning, ICML '99, pp. 124–133. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999)

8. Freund, Y., Schapire, R.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences **55**(1), 119–139 (1997)
9. García, C., Marín, H.: An integrated model to study the structural and financial sustainability of agricultural enterprises. Acta Horticulturae **674**, 313–320 (2005)
10. Gutiérrez, P.A., Hervás-Martínez, C., Martínez-Estudillo, F.J.: Logistic regression by means of evolutionary radial basis function neural networks. IEEE Transactions on Neural Networks **22**(2), 246–263 (2011). Submitted
11. Gutiérrez, P., Segovia-Vargas, M., Salcedo-Sanz, S., Hervás-Martínez, C., Sanchis, A., Portilla-Figueras, J., Fernández-Navarro, F.: Hybridizing logistic regression with product unit and RBF networks for accurate detection and prediction of banking crises. Omega **38**(5), 333–344 (2010). Cited By 7
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. Special Interest Group on Knowledge Discovery and Data Mining Explorer Newsletter **11**, 10–18 (2009)
13. Haykin, S.: Neural Networks: A comprehensive Foundation. Prentice Hall, 3rd edition (2008)
14. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis Journal **6**(5) (2002)
15. Johnson, R.A., Wichern, D.W. (eds.): Applied Multivariate Statistical Analysis. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1988)
16. Kibler, D.F., Aha, D.W., Albert, M.K.: Instance-based prediction of real-valued attributes. Computational Intelligence **5**, 51 (1989)
17. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Machine Learning **59**(1-2), 161–205 (2005)
18. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences **250**, 113 – 141 (2013)
19. MacFarlane, R.: Modelling the interaction of economic and socio-behavioural factors in the prediction of farm adjustment. Journal of Rural Studies **12**(4), 365–374 (1996)
20. Martínez-Estudillo, A.C., Martínez-Estudillo, F.J., Hervás-Martínez, C., García, N.: Evolutionary product unit based neural networks for regression. Neural Networks **19**(4), 477–486 (2006)
21. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
22. Sadras, V., Roget, D., Krause, M.: Dynamic cropping strategies for risk management in dry-land farming systems. Agricultural Systems **76**(3), 929–948 (2003)
23. Torres, M., Hervás, C., García, C.: Multinomial logistic regression and product unit neural network models: Application of a new hybrid methodology for solving a classification problem in the livestock sector. Expert Systems with Applications **36**(10), 12,225–12,235 (2009)
24. Webster, J.: Assessing the economic consequences of sustainability in agriculture. Agriculture, Ecosystems and Environment **64**(2), 95–102 (1997)