Neural Network-Based Approach for Supervised Nonlinear Feature Selection

Mamadou Kanouté, Edith Grall-Maës and Pierre Beauseroy
Computer Science and Digital Society Laboratory (LIST3N), Université de Technologie de Troyes, Troyes, France

Keywords: Neural Network, Multi-Output Regression, Supervised Nonlinear Feature Selection.

Abstract: In machine learning, the complexity of training a model increases with the size of the considered feature space. To overcome this issue, feature or variable selection methods can be used for selecting a subset of relevant variables. In this paper we start from an approach initially proposed for classification problems based on a neural network with one hidden layer in which a regularization term is incorporated for variable selection and then show its effectiveness for regression problems. As a contribution, we propose an extension of this approach in the multi-output regression framework. Experiments on synthetic data and real data show the effectiveness of this approach in the supervised framework and compared to some methods of the literature.

1 INTRODUCTION

The latest technological advances allow the collection of data from various devices . They can produce many measurements of different types (categorical, continuous) allowing them to describe the monitored system. To infer some results not all features might be useful, some contain no information or are redundant. To set up a model on these data for a prediction problem, for example, these variables must be studied to keep only the relevant ones. Variable selection is a data analysis technique that allows the selection of relevant variables by removing redundancy and non-informative variables and the selection is made with respect to one or more target variables. These target variables can be categorical or continuous. Many methods of variable selection have been proposed for the case of a single target variable using statistical methods, information theory, and neural networks and can be categorized into three groups:

- Filter methods use statistical measures between the target variable and other variables to select important variables such as (He et al., 2005) where the laplacian score is used as a statistical measure.
- Wrapper methods are based on learning models whose relevance of the selected variables depends on the performance of the learning model, in (Maldonado and Weber, 2009) the authors do the feature selection using Support Vector Machine as a learning model.
- Embedded methods add the selection constraint in the initial formulation of the prediction model as a regularization term to properly estimate the target variable while determining the important ones. One of the best-known methods is Lasso (Tibshirani, 2011), an approach that adds a regularization l_1 in the formulation of a linear prediction problem to constrain weights to be sparse coefficients representing the predictor variables.

Many of these methods exploit only the linear relationships between the variables. In (Yamada et al., 2014), (Song et al., 2012), (Song et al., 2007) the authors propose a nonlinear feature selection method for a single target variable based on Hilbert-Schmidt Independence Criterion (Gretton et al., 2005), a nonlinear dependency measure using kernel methods. The complexity of this approach lies in finding the right kernel and its parameter.

In recent years, another type of variable selection methods in the supervised framework has attracted the attention of researchers. It uses several target variables based on multi-task learning (Zhang and Yang, 2018), a subdomain of machine learning in which several learning tasks are solved at the same time while exploiting commonalities and differences between the tasks. A good example is multi-output regression, a regression problem with several continuous target variables as tasks. Many applications for multi-output regression have been studied. Approaches in the linear and nonlinear case have been proposed, in particular those using single hidden layer neural networks.

In this paper, we are interested in problems of nonlinear supervised variable selection with one or several target variables applied to regression problems on continuous variables. Starting from a variable se-

Kanouté, M., Grall-Maës, E. and Beauseroy, P.
Neural Network-Based Approach for Supervised Nonlinear Feature Selection.
DOI: 10.5220/0012185700003595
In *Proceedings of the 15th International Joint Conference on Computational Intelligence (IJCCI 2023)*, pages 431-439
ISBN: 978-989-758-674-3; ISSN: 2184-3236

lection approach initially used for classification problems with a single target variable, our contribution is as follows:

- Apply this method for regression problems with a single target variable.
- Propose an extension of this approach in the case of selection with several target variables.

The core part of the paper is organized as follows: in section 2, notations are introduced and related works are detailed. In section 3, the method used as well as its extension in the multi-output regression is exposed. Experimental results are given and discussed in section 4. Finally, in section 5 conclusions are drawn and perspectives are proposed.

2 NOTATIONS AND RELATED WORKS

In this section, firstly some notations used in the paper are given and secondly an overview of previous work related to our work is given.

2.1 Notations

The following notations are used:

- $\mathcal{S}$ is the set of variables in the dataset.
- $\mathcal{S}_Y$ and $\mathcal{S}_X$ form a partition of $\mathcal{S}$. They denote respectively the set of target variables and the set of predictor variables.
  $\mathcal{S}_X \cup \mathcal{S}_Y = \mathcal{S}$ and $\mathcal{S}_X \cap \mathcal{S}_Y = \emptyset$.
- X and Y are matrices of observations whose variables are respectively in $\mathcal{S}_X$ and $\mathcal{S}_Y$.
- For any matrix M, the vectors M_i and M^j are the i^{th} row and j^{th} column of M respectively.
- For any matrix $M \in \mathbb{R}^{n\times d}$, the Frobenius norm (Noble and Daniel, 1997) is defined as follows:

$$||M||_F = \sqrt{tr(M^T M)} = \sqrt{\sum_{1\leq i\leq n}\sum_{1\leq j\leq d} m_{ij}^2} \quad (1)$$

- For any matrix $M \in \mathbb{R}^{n\times d}$, the $l_{2,1}$ norm (Ding et al., 2006) is defined as follows:

$$||M||_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d} m_{ij}^2} \quad (2)$$

  The $||.||_{2,1}$ norm applies the l_2 norm to the column elements and the l_1 norm to the computed row norm. This norm, therefore, makes it possible to impose sparsity on the rows of M.

- For two matrices $M \in \mathbb{R}^{n\times d}$ and $\widehat{M} \in \mathbb{R}^{n\times d}$, the Mean Squared Error (MSE) is defined as follows:

$$\begin{aligned} MSE(M,\widehat{M}) &= \frac{1}{nd}\sum_{j=1}^{d}\sum_{i=1}^{n}(m_{ij}-\widehat{m}_{ij})^2 \\ &= \frac{1}{nd}||M-\widehat{M}||_F^2 \end{aligned} \quad (3)$$

2.2 Related Works

In this section, some selection methods related to our work are described. In (Obozinski et al., 2006) the authors propose Multi-task Lasso. It is a selection approach based on multi-task learning, a concept allowing to jointly solve several tasks defined by a set of features $\mathcal{S}_Y$ and regularization $l_{2,1}$ defined in section 2.1. Multi-task Lasso therefore makes it possible to jointly solve several related regression tasks i.e the variables of interest in $\mathcal{S}_Y$ by simultaneously selecting variables in $\mathcal{S}_X$ common to the different tasks. The regression coefficient matrix of variables noted W is determined by minimizing the following expression:

$$\mathcal{L}_C(W) = ||Y-XW||_F^2 + C||W||_{2,1}, \quad (4)$$

C is the regularization parameter for sparsity. The larger is C the sparser is W. This parameter tunes the trade-off between the estimation of the target variables and the number of selected variables. Once C^* the optimal C has been determined according to a criterion, the importance of each variable is determined by calculating the Euclidean norm of its corresponding row in W, and variables with low impact can be removed from the model. This method exploits only the linear relationships between the variables. In (Wang et al., 2021), the authors propose NFSN (Nonlinear Feature Selective Networks), a nonlinear approach for variable selection for several target variables. This method is based on a single hidden layer neural network and the addition of a regularization $l_{2,1}$ on the weight matrix of the hidden layer for joint selection. The expression to be optimized is:

$$\mathcal{L}_C(\Theta) = \frac{1}{2N}||Y-\widehat{Y}||_2^2 + C||W^{(1)}||_{2,1} \quad (5)$$

where

- $\Theta = \{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$ is the set of neural network parameters to be optimized where $W^{(1)}$ and $W^{(2)}$ are respectively the weight matrices of the hidden layer and the output layer and $b^{(1)}$ and $b^{(2)}$ are the corresponding biases.
- $\widehat{Y} = \sigma_1(XW^{(1)} + b^{(1)})W^{(2)} + b^{(2)}$ where σ_1 is an activation function.
- C is the regularization parameter for sparsity (as defined in Equation 4).

- N is the sample size.

Once C^* is determined, the importance of each variable i is determined by calculating the Euclidean norm of its corresponding row in $W^{(1)}$ i.e $||W_i^{(1)}||_2$.

These two methods allow variable selection with several target variables, Multi-task Lasso exploits only linear relationships between variables while NFSN exploits nonlinear relationships between variables. In both of these approaches, the importance of variables is determined by taking the Euclidean norm over the rows of a weight matrix and ranking the variables according to these calculated norms. Here the goal is to determine in a nonlinear way the coefficient associated with each variable.

3 PROPOSED APPROACH

In this part, the formulation of the based approach developed is introduced. This approach was initially introduced to tackle a classification problem. An extension of this approach is proposed for the multi-output regression framework. In section 4 it is shown that it can be used for regression problems.

3.1 Initial Approach for Classification

In (Challita et al., 2016) the variable selection is performed using a type of neural network called Extreme learning machine (Huang et al., 2006) . It is a neural network with one single hidden layer where the weight matrix of the hidden layer is randomly generated and not updated. Only the weight matrix of the output layer is updated. According to (Huang et al., 2006) these models can produce good generalization performance and have a much faster learning process than neural networks trained using gradient backpropagation. The variable selection method is based on the idea of assigning a weight to each attribute. In the beginning, the weights of all attributes are equal. The main goal of the method is to adjust the weights of the different attributes to minimize the classification error. Attributes with high values of weights are important and should be kept. Attributes with low values of weights are not important and can be removed. An illustration of the approach is given in Figure 1.
Let N be the sample size and p the number of variables.
Let $NNeur$ be the number of neurons in the hidden layer.
Let $X = [a_1, \cdots, a_p]^T \in \mathbb{R}^{p \times N}$ where $a_i \in \mathbb{R}^N$ is the realisation of feature i for all observations and Y is a vector of labels containing -1 or 1.

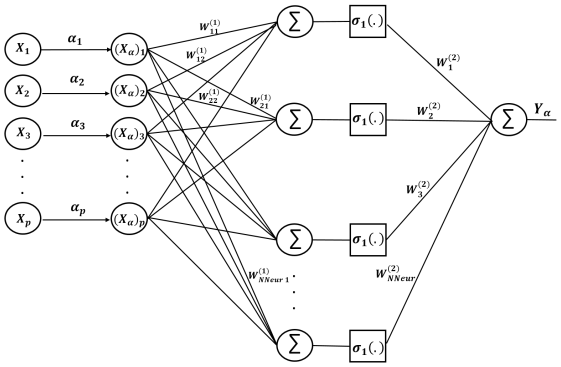


Figure 1: Architecture of the used approach.

The selection of features is done by minimizing

$$\mathcal{L}_{\lambda,C}(\Theta) = ||Y - Y_\alpha||_2^2 + \lambda||W^{(2)}||_2^2 + C\sum_{i=1}^{p}(D_\alpha)_{ii} \quad (6)$$

where

- $Y_\alpha \in \mathbb{R}^{N\times 1}$ is the network output. It is defined as follows

$$Y_\alpha = S_\alpha W^{(2)} = \sigma_1[(W^{(1)}X_\alpha)^T]W^{(2)} \quad (7)$$

  where
  - σ_1 is an activation function. In (Challita et al., 2016), $\sigma_1(.) = tanh(.)$.
  - $W^{(1)} \in \mathbb{R}^{NNeur\times(p+1)}$ is the weight matrix of the hidden layer that contains the bias coefficient. It is a random matrix.
  - $W^{(2)} \in \mathbb{R}^{NNeur\times 1}$ is the weight matrix of the network output also containing the bias.
  - $X_\alpha = D_\alpha X'$ is a $(p+1)\times N$ matrix whose variables are weighted
    where
    * $X' = \begin{pmatrix} X \\ \mathbf{1}_N^T \end{pmatrix}$ is $(p+1)\times N$ matrix where $\mathbf{1}_N$ is a vector of $\mathbb{R}^N$ containing only 1.
    * $D_\alpha \in \mathbb{R}^{(p+1)\times(p+1)}$ is a diagonal matrix containing the weight associated with each variable such that $(D_\alpha)_{i,i} = \alpha_i$
      where
      $\alpha_i \in [0,1]$ is the weight associated with each variable i for $i = 1, \cdots, p$.
      α_{p+1} is the weight associated with the fixed input (bias). $\alpha_{p+1} = 1$.
- C is the regularization parameter for sparsity that allows setting some α_i to 0.

- λ is the regularization parameter allowing better stability and better generalization.

$W^{(2)}$ and D_α are the unknowns, $\Theta = (W^{(2)}, D_\alpha)$.

3.2 Determination of Parameters

The determination of the optimal parameters $W^{(2)*}$ and D_α^* is crucial for estimating the target variable and the selection of variables. To optimize the model, $W^{(2)}$ and D_α are updated alternately and iteratively. That is, $W^{(2)}$ is updated with D_α fixed and vice versa. D_α is initialized as an identity matrix.

For fixed D_α, in (Challita et al., 2016) $W^{(2)}$ is updated by calculating the derivative of Equation 6 with respect to $W^{(2)}$, which leads to the simple closed form solution:

$$W^{(2)*} = \frac{S_\alpha^T Y}{(S_\alpha S_\alpha^T + \lambda I)} \tag{8}$$

For fixed $W^{(2)}$, D_α the diagonal matrix with α_i as its diagonal entries is updated. To take into account the constraints on α_i for $i = 1, \ldots, p$ defined in section 3.1 , the optimization problem is reformulated as follows:

$$\begin{aligned} &\underset{\alpha_i}{\text{minimize}} \quad \mathcal{L}_{\lambda,C}(\Theta) \\ &\text{subject to} \quad \alpha_i - 1 \leq 0, -\alpha_i \leq 0,\ i = 1, \ldots, p. \end{aligned} \tag{9}$$

As in (Challita et al., 2016), the partial derivative of Equation 6 with respect to α_i is approximated by numerical methods. The optimization problem of Equation 9 is solved by optimization algorithms.

3.3 Multi-Output Regression

Based on the formulation given in (Challita et al., 2016) which is adapted to the single task problem, a new method named FS-ELM meaning Feature Selection using Extreme Learning Machine is proposed. This method can tackle multi output regression problems where the number of variables in $\mathcal{S}_Y$ is greater than 1. The proposed method replaces the l_2 norm in the objective function and the constraint on $W^{(2)}$ by the Frobenius norm where $W^{(2)} \in \mathbb{R}^{NNeur \times card(\mathcal{S}_Y)}$. $\mathcal{L}_{\lambda,C}(\Theta)$ is reformulated as follows:

$$\mathcal{L}_{\lambda,C}(\Theta) = ||Y - Y_\alpha||_F^2 + \lambda ||W^{(2)}||_F^2 + C \sum_{i=1}^{p} (D_\alpha)_{ii} \tag{10}$$

where

- The derivative of $\mathcal{L}(\Theta)$ with respect to $W^{(2)}$ remains the same as in Equation 8.
- The update of D_α remains the same as defined in Equation 9.

4 EXPERIMENTS

In this part, two sets of variables $\mathcal{S}_X$ and $\mathcal{S}_Y$ corresponding respectively to available variables and variables to be inferred are assumed to be defined. The effectiveness of the proposed approach to select a subset of relevant variables in $\mathcal{S}_X$ for estimating $\mathcal{S}_Y$ is shown. The subsection 4.1 is about the evaluation of FS-ELM on synthetic data and the subsection 4.2 on real-world data. For a single target variable, the proposed method is compared with Lasso, NFSN, and for several target variables, the comparison is made with Multi-task Lasso, NFSN. To show the effectiveness of the proposed method and to compare it with other methods, the procedure is composed of three steps:

- Determine C^* and λ^* the optimal values of C and λ according to a criterion on the MSE.
  - for $C \in I_C$ with $I_C = \{10^{-4}, 10^{-3}, \ldots, 10^3, 10^4\}$
    - for $\lambda \in I_\lambda$ with $I_\lambda = \{10^{-4}, 10^{-3}, \ldots, 10^3, 10^4\}$
      - Compute $\widehat{Y}^{(\lambda,C)}$ the estimate of Y associated with C and λ on a train data set.
  - Choose $(C^*, \lambda^*) \in I_C \times I_\lambda$ such that

$$(C^*, \lambda^*) = \underset{(C,\lambda) \in I_C \times I_\lambda}{\operatorname{argmin}} \ MSE(Y, Y^{(\lambda,C)}) \tag{11}$$

  $(C^*, \lambda^*) \in I_C \times I_\lambda$ is a pair of values that minimize $MSE(Y, Y^{(\lambda,C)})\ \forall\ (C, \lambda) \in I_C \times I_\lambda$ using a test data set.

  For Multi-task Lasso, NFSN and Lasso approaches where C is the only parameter to be tuned, the procedure is similar to the one above but only C^* is determined.

- Once hyperparameters are chosen, for each approach, rank the variables according to their importance.
  - For Lasso, rank the variables according to the ordered values of the absolute value of the coefficients of the linear regression model.
  - For Multi-task Lasso and NFSN, rank the variables as defined in section 2.2.
  - For FS-ELM, rank the variables according to the scaling factors α_i.
- Evaluate the pertinence of ranking for each approach by building p models on the train data set and evaluating them on the test data set by keeping from 1 to p variables corresponding to the highest rank.

  The evaluation model used is a single hidden layer neural network with 500 neurons. The activation function is relu and the optimizer is

adam.
The relevance of the selected variables is evaluated using the MSE of the estimated model.

To avoid scaling problems, the variables of matrices X and Y are normalized (a pre-processing technique of removing the mean and scaling to unit variance applied to each variable).
The validation of the results is done by 5-fold cross-validation.

4.1 Synthetic Data Set

Firstly, 8 Gaussian features were defined. Then 10 random features that depend on these 8 Gaussian features with nonlinear relationships were defined. Finally, 4 other independent Gaussian features were added. Then data were generated from these 22 features as described below.

- $f_1 \sim \mathcal{N}(1,5^2)$; $f_2 \sim \mathcal{N}(0,2^2)$; $f_3 \sim \mathcal{N}(2,7^2)$
- $f_4 \sim \mathcal{N}(5,3^2)$, $f_5 \sim \mathcal{N}(0,1)$; $f_6 \sim \mathcal{N}(0,0.3^2)$
- $f_7 \sim \mathcal{N}(3,2^2)$; $f_8 \sim \mathcal{N}(11,1)$
- $f_9 = sin(f_1) + \varepsilon_{f_9}, \varepsilon_{f_9} \sim \mathcal{N}(0,0.08^2)$
- $f_{10} = log(|f_3|) + \varepsilon_{f_{10}}, \varepsilon_{f_{10}} \sim \mathcal{N}(0,0.08^2)$
- $f_{11} = cos(f_2) + \varepsilon_{f_{11}}, \varepsilon_{f_{11}} \sim \mathcal{N}(0,0.1^2)$
- $f_{12} = f_1^2 sin(\frac{f_3}{f_1}) + \varepsilon_{f_{12}}, \varepsilon_{f_{12}} \sim \mathcal{N}(0,0.04^2)$
- $f_{13} = f_2^3 + f_2 f_3^2 - f_2^2 + f_3^2 + \varepsilon_{f_{13}}$, $\varepsilon_{f_{13}} \sim \mathcal{N}(0,0.02^2)$
- $f_{14} = f_5(f_4^2 + log(f_5^2)) + \varepsilon_{f_{14}}, \varepsilon_{f_{14}} \sim \mathcal{N}(0,0.08^2)$
- $f_{15} = f_5^2 + f_6(sin(f_5 + f_6^2)) + \varepsilon_{f_{15}}$, $\varepsilon_{f_{15}} \sim \mathcal{N}(0,0.08^2)$
- $f_{16} = \frac{f_7 f_8}{f_7^2 + f_8^2} + \varepsilon_{f_{16}}, \varepsilon_{f_{16}} \sim \mathcal{N}(0,0.04^2)$
- $f_{17} = sin(e^{-f_7^2}) + \varepsilon_{f_{17}}, \varepsilon_{f_{17}} \sim \mathcal{N}(0,0.01^2)$
- $f_{18} = cos(sin(f_8)) + \varepsilon_{f_{18}}, \varepsilon_{f_{18}} \sim \mathcal{N}(0,0.08^2)$
- $f_{19}, f_{20}, f_{21}, f_{22} \sim \mathcal{N}(0,1)$.

The set of variables is $\mathcal{S} = \{f_1, \cdots, f_{22}\}$

On the generated data, experiments have been done in two cases for $\mathcal{S}_Y$, with $\mathcal{S}_X = \mathcal{S} \setminus \mathcal{S}_Y$.

For the first experiment, $\mathcal{S}_Y = \{f_{15}\}$. Figure 2a shows the estimated value of the mean of the MSE for the variables of $\mathcal{S}_Y$ versus $log(C)$. For Lasso $C^* = 10^{-2}$, for FS-ELM $C^* = 10^2, \lambda^* = 10^{-3}$ and for NFSN $C^* = 10^{-4}$. After the choice of the regularization parameters, the list of ranked variables for Lasso, FS-ELM and NFSN is given in Table 1. Figure 3a shows the estimated value of the mean of MSE versus the number of most important variables used to build the model. It may be noticed that for NFSN and FS-ELM the first 2 most important variables allow to estimate well the target variable while for Multi-task Lasso it takes the first 14 most important variables. As the same variables were selected by NFSN and FS-ELM, the green curve is exactly underneath the red curve.

Table 1: List of ranked variables for each approach. Variables in bold are the ideal variables that should be selected by the approaches.

(a) $\mathcal{S}_Y = \{f_{15}\}$.

Lasso	FS-ELM	NFSN
f_{14}	$\mathbf{f_5}$	$\mathbf{f_5}$
f_{22}	$\mathbf{f_6}$	$\mathbf{f_6}$
f_7	f_{18}	f_{14}
$\mathbf{f_6}$	f_{21}	f_{18}
f_4	f_{16}	f_{22}
f_{13}	f_8	f_{19}
f_3	f_{22}	f_{10}
f_{10}	f_3	f_{11}
f_{21}	f_4	f_8
f_{20}	f_{11}	f_2
f_9	f_1	f_1
f_1	f_{20}	f_4
f_{17}	f_2	f_9
$\mathbf{f_5}$	f_9	f_{20}
f_{12}	f_{12}	f_{17}
f_{11}	f_{13}	f_{13}
f_{19}	f_{17}	f_{12}
f_{18}	f_7	f_7
f_2	f_{14}	f_3
f_{16}	f_{19}	f_{16}
f_8	f_{10}	f_{21}

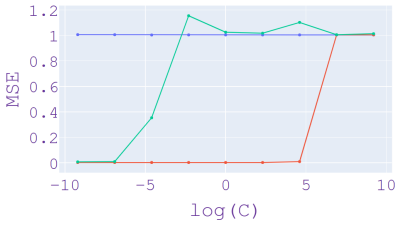
(b) $\mathcal{S}_Y = \{f_{11}, f_{17}, f_{18}\}$.

Multi-task Lasso	FS-ELM	NFSN
f_{16}	$\mathbf{f_7}$	$\mathbf{f_7}$
$\mathbf{f_7}$	$\mathbf{f_2}$	$\mathbf{f_8}$
$\mathbf{f_8}$	$\mathbf{f_8}$	$\mathbf{f_2}$
f_3	f_{16}	f_{16}
f_5	f_{22}	f_{19}
$\mathbf{f_2}$	f_3	f_3
f_6	f_1	f_{15}
f_{13}	f_{14}	f_9
f_9	f_5	f_{20}
f_{14}	f_{21}	f_{22}
f_{22}	f_6	f_{13}
f_{20}	f_{10}	f_1
f_{12}	f_{20}	f_{14}
f_{21}	f_4	f_6
f_{15}	f_{13}	f_5
f_1	f_{15}	f_4
f_{10}	f_9	f_{10}
f_4	f_{19}	f_{12}
f_{19}	f_{12}	f_{21}

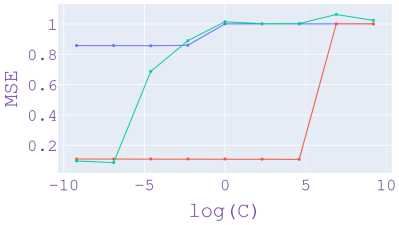
For the second experiment, $\mathcal{S}_Y = \{f_{11}, f_{17}, f_{18}\}$. Figure 2b shows the estimated value of the mean of the MSE for the variables of $\mathcal{S}_Y$ versus $log(C)$. For Multi-task Lasso $C^* = 10^{-2}$, for FS-ELM $C^* = 10^2, \lambda^* = 10^{-1}$, for NFSN $C^* = 10^{-3}$. Table 1b contains the list of ranked variables for Multi-task Lasso, FS-ELM and NFSN after the choice of regularization parameters. Figure 3b shows the estimated value of the mean of MSE versus the number of most important variables taken. It may be noticed that for NFSN and FS-ELM the first 3 most important variables allow to estimate well the target variable while for Lasso it takes the first 6 most important variables.

4.2 Real-World Data Sets

In this part, the proposed method is evaluated on the real-world data sets and compared to other methods for one and several target variables. Table 2 contains the list of real-world data used as well as the number of target variables, number of predictor variables, and number of samples. Some information about the data sets in Table 2 as well as the pre-processing of data

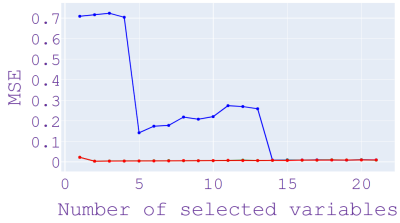


(a) $\mathcal{S}_Y = \{f_{15}\}$ and $\mathcal{S}_X = \mathcal{S} \setminus \mathcal{S}_Y$.

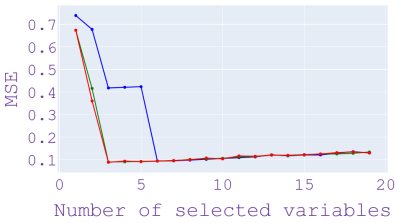


(b) $\mathcal{S}_Y = \{f_{11}, f_{17}, f_{18}\}$ and $\mathcal{S}_X = \mathcal{S} \setminus \mathcal{S}_Y$.

Figure 2: MSE versus $log(C)$ on synthetic data. Lasso (blue), NFSN (green), FS-ELM (red).

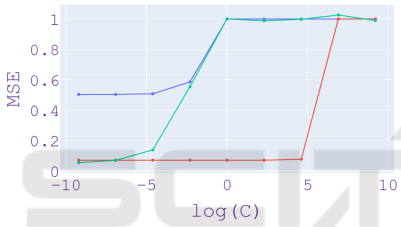


(a) $\mathcal{S}_Y = \{f_{15}\}$ and $\mathcal{S}_X = \mathcal{S} \setminus \mathcal{S}_Y$.

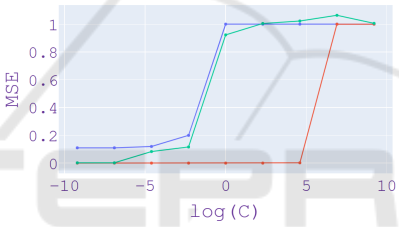


(b) $\mathcal{S}_Y = \{f_{11}, f_{17}, f_{18}\}$ and $\mathcal{S}_X = \mathcal{S} \setminus \mathcal{S}_Y$.

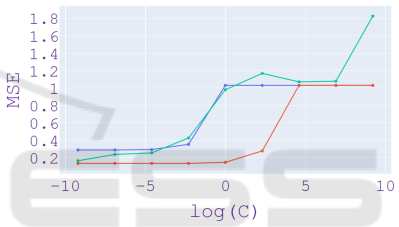
Figure 3: MSE versus number of most important variables on synthetic data. Lasso (blue), NFSN (green), FS-ELM (red).



(a) Bike sharing data set.

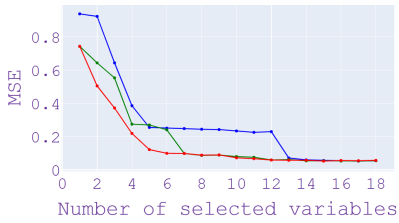


(b) Air quality data set.

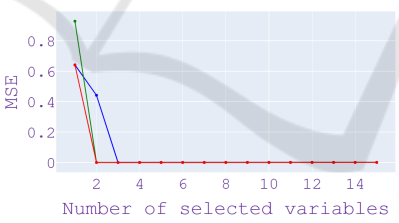


(c) Boston house data set.

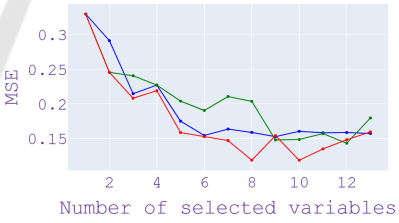
Figure 4: MSE versus $log(C)$ on real-world data with a single target. Lasso (blue), NFSN (green), FS-ELM (red).



(a) Bike sharing dataset.

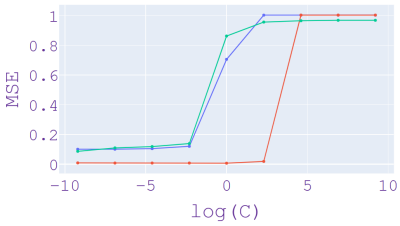


(b) Air quality data set.

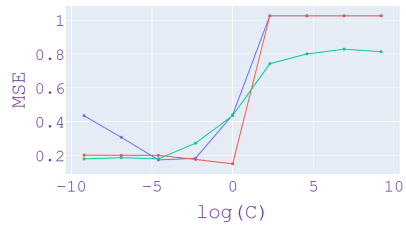


(c) Boston house data set.

Figure 5: MSE versus number of important variables to keep for selection with a single target variable on real-world data sets. Lasso (blue), NFSN (green), FS-ELM (red).

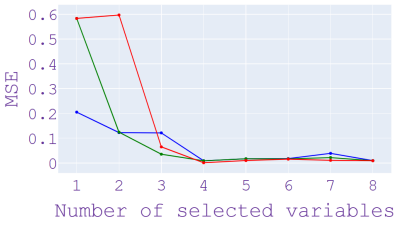


(a) Enb data set.

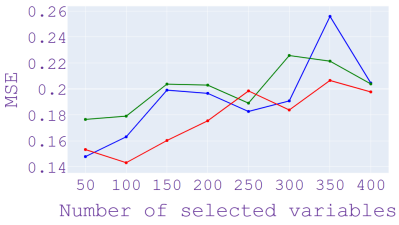


(b) Atp1d data set.

Figure 6: MSE versus $log(C)$ on real-world data sets with several target variables. Multi-task Lasso (blue), NFSN (green), FS-ELM (red).



(a) Enb data set.



(b) Atp1d data set.

Figure 7: MSE versus the number of most important variables on real-world data sets with several target variables. Multi-task Lasso (blue), NFSN (green), FS-ELM (red).

done are described below.

- Bike sharing dataset
  This data set contains monitoring data of rental bike users in a city with 16 variables including one target variable and 15 predictor variables. Only the file "hour.csv" on UCI website is used in this paper. Among the 15 predictors variables, 7 variables are continuous, 3 variables are binary categorical and 5 are cyclical discrete variables. The sine and cosine transformation is applied to the cyclic variables, they are then removed and the continuous variables are normalized. The selection approaches are applied to the obtained 20 variables in order to estimate the target variable.
- Air quality dataset
  This data contains the responses of a gas multisensor device deployed on the field in an Italian city. It is composed of 15 variables including a target variable and 14 predictors variables including 12 continuous variables, a time variable and a date variable. A new variable called month is deduced from the variable date and the variable called hour is deduced from variable time. The sine and cosine transformation is applied to the cyclic variables month and hour, they are then removed. The selection approaches are applied to 17 variables for estimating the target variable.
- Boston house
  This data set contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. There are 14 variables including one target variable and 13 continuous predictor variables.
- Enb dataset
  This data set of 10 variables includes 2 target variables the heating load and cooling load requirements of the building and 8 continuous predictor variables such as glazing area, roof area, and overall height, ... The data set is taken from Mulan, an open-source Java library for learning from multi-label data sets and it can also be downloaded from their github https://github.com/tsoumakas/mulan.

Table 2: Real-world data sets for variable selection.

Name	Size	Features	Targets	Source
Bike sharing	17 389	15	1	(Fanaee-T and Gama, 2013)
Air quality	9 357	14	1	(De Vito et al., 2008)
Boston house	506	13	1	(Harrison and Rubinfeld, 1978)
Enb	768	8	2	(Tsanas and Xifara, 2012)
Atp1d	337	411	6	(Xioufis et al., 2012)

- Atp1d
  This dataset of 337 observations is about the prediction of airline ticket prices. There are 417 variables including 6 variables as targets and 411 predictor variables. The data set is taken from Mulan.

The cases with one target are first tackled. Figure 4 shows the estimated value of the mean of the MSE between Y and its estimate versus $log(C)$ by 5-fold cross-validation for each approach on the Bike sharing, Air quality, and Boston data sets. It can be noticed the stability of performance of FS-ELM for large values of C compared to other methods.
For each approach and each data set, the chosen C^* is described below:

- On Bike sharing dataset, $C^* = 10^{-4}$ for Lasso, $C^* = 10, \lambda^* = 10^{-4}$ for FS-ELM and $C^* = 10^{-4}$ for NFSN.
- On Air quality dataset, $C^* = 10^{-4}$ for Lasso, $C^* = 10^{-1}, \lambda^* = 10^{-1}$ for FS-ELM and $C^* = 10^{-4}$ for NFSN.
- On Boston house dataset, $C^* = 10^{-4}$ for Lasso, $C^* = 10^{-1}, \lambda^* = 10^{-2}$ for FS-ELM and $C^* = 10^{-4}$ for NFSN.

Once the regularization parameters have been determined for each approach, the most important variables are taken gradually, then an estimate is made to assess the relevance of the variables taken and to determine the number of variables to keep. The number of important variables taken successively is $\{1,2,\ldots,20\}$ on Bike sharing data set, $\{1,2,\ldots,17\}$ on Air quality data set and $\{1,2,\ldots,13\}$ on Boston house data set. Figure 5 shows the MSE between Y and its estimate versus the number of important variables taken successively for each approach on Bike

sharing, Air quality, Boston house data sets. It can be noticed that in general FS-ELM manages to select the variables better compared to the other approaches. Precisely:

- On Bike sharing data set, FS-ELM performs well in the variable selection compared to Lasso and NFSN.
- On Air quality data set, the two first important variables selected by FS-ELM and NFSN can estimate well the target variable.
- On the Boston house data set, FS-ELM performs well compared to NFSN. Indeed, FS-ELM has the minimum MSE for any number of selected variables. There are some variances in the MSE because there are only 508 samples.

Figure 6 shows the estimated value of the mean of the MSE by 5-fold cross-validation between Y and its estimate versus $log(C)$ for each approach on the data sets with several target variables. It can be noticed that FS-ELM has greater stability for regularization parameters than the other methods. For each approach and each data set, the chosen C^* is described below:

- On Enb data set, $C^* = 10^{-4}$ for Multi-task Lasso, $C^* = 1, \lambda^* = 10^{-3}$ for FS-ELM, $C^* = 10^{-4}$ for NFSN.
- On Atp1d data set, $C^* = 10^{-2}$ for Multi-task Lasso, $C^* = 1, \lambda^* = 10^{-2}$ for FS-ELM, $C^* = 10^{-2}$ for NFSN.

Once the regularization parameters have been determined for each approach, the variables are ranked for each approach. The number of important variables taken successively is $\{1, 2, \ldots, 8\}$ on Enb data set and $\{50, 100, 150, \ldots, 400\}$ on Atp1d data set. Figure 7 shows the estimated value of the mean of the MSE by 5-fold cross-validation between Y and its estimate versus the number of important variables taken successively for each approach and on the data sets Enb, Atp1d. It can be noticed that in general, FS-ELM manages to select well the relevant variables and reach the best performance with Atp1d which is the most challenging case.

The proposed method successfully selects the relevant variables on regression problems for one and several target variables. In addition, it can be noticed that generally, FS-ELM selects better compared to NFSN and Multi-task Lasso.

5 CONCLUSIONS

In this paper, starting from an approach that was initially proposed for a classification problem with a single target variable, we first showed its feasibility for regression problems with a single target variable, then proposed an extension in the framework of multi-output regression for variable selection with several target variables. Finally, many experiments made on synthetic data and real data confirm the effectiveness of the proposed approach.
The future works would be to:

- Calculate the partial derivative of $\mathcal{L}_{\lambda,C}(\Theta)$ with respect to the α_i since it was not calculated in the initial formulation and this work to improve the optimization algorithm.
- Propose an approximation of the matrix division made in Equation 8 to reduce the complexity of the optimization.
- Apply the proposed extension to the unsupervised nonlinear variable selection problems for continuous variables.

ACKNOWLEDGEMENT

This work was supported by Labcom-DiTeX, a joint research group in Textile Data Innovation between Institut Français du Textile et de l'Habillement (IFTH) and Université de Technologie de Troyes (UTT).

REFERENCES

Challita, N., Khalil, M., and Beauseroy, P. (2016). New feature selection method based on neural network and machine learning.

De Vito, S., Massera, E., Piga, M., Martinotto, L., and Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B Chemical*.

Ding, C., Zhou, D., He, X., and Zha, H. (2006). R 1-pca: Rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*.

Fanaee-T, H. and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms.

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*.

He, X., Cai, D., and Niyogi, P. (2005). Laplacian score for feature selection.

Huang, G., Zhu, Q.-Y., and Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*.

Maldonado, S. and Weber, R. (2009). Weber, r.: A wrapper method for feature selection using support vector machines. inf. sci. 179(13), 2208-2217.

Noble, B. and Daniel, J. W. (1997). Applied linear algebra. 2nd ed.

Obozinski, G., Taskar, B., and Jordan, M. (2006). Multitask feature selection.

Song, L., Bedo, J., Borgwardt, K. M., Gretton, A., and Smola, A. (2007). Gene selection via the bahsic family of algorithms. volume 23, pages i490–i498. Bioinformatics.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature selection via dependence maximization. JMLR.org.

Tibshirani, R. (2011). Regression shrinkage selection via the lasso. In *Journal of the Royal Statistical Society Series B*.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*.

Wang, Z., Nie, F., Zhang, C., Wang, R., and Li, X. (2021). Joint nonlinear feature selection and continuous values regression network.

Xioufis, E. S., Groves, W., Tsoumakas, G., and Vlahavas, I. P. (2012). Multi-label classification methods for multi-target regression. *CoRR*.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. MIT Press - publishers.

Zhang, Y. and Yang, Q. (2018). An overview of multi-task learning.