Unsupervised Feature Selection Using Extreme Learning Machine

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

Keywords: Neural Network, Sparse Learning, Nonlinear Method, Unsupervised Feature Selection.

Abstract: In machine learning, feature selection is an important step in building an inference model with good generalization capacity when the number of variables is large. It can be supervised when the goal is to select features with respect to one or several target variables or unsupervised where no target variable is considered and the goal is to reduce the number of variables by removing redundant variables or noise. In this paper, we propose an unsupervised feature selection approach based on a model that uses a neural network with a single hidden layer in which a regularization term is incorporated to deal with nonlinear feature selection for multi-target regression problems. Experiments on synthetic and real-world data and comparisons with some methods in the literature show the effectiveness of this approach in the unsupervised framework.

1 INTRODUCTION

Nowadays with technological advances (storage and capturing systems), data can be collected in different ways in which variables can be numerous and of different types (continuous or categorical). These variables can be used to infer some results or explain certain relationships or trends. However, some of them can be not informative or redundant and must be removed to reduce the cost of data storage or create less complex and interpretable models. Variable selection is a machine learning technique that determines a subset of relevant variables from an original set. The selection can be supervised or unsupervised. The supervised framework allows the selection of relevant variables with respect to one or several target variables. The unsupervised framework that concerns our work allows to perform the selection without target variables; the aim is to reduce the redundancy within the variables or to select them while preserving the geometric structure of the data. Many methods have been proposed for variable selection in the unsupervised setting and can be categorized into 3 classes (Solorio-Fernández et al., 2020).

- Filter methods use statistical measures between variables to select important variables based on intrinsic properties of the data such as (He et al., 2005) where the Laplacian score is used as a statistical measure to determine important variables.
- Wrapper methods are methods based on the performance of a learning algorithm. Many of these methods in the unsupervised setting are based on clustering algorithms and the relevance of the selected variables depends on their contribution to the results of the clustering . In (Cai et al., 2010), authors propose Multi-Cluster Feature Selection (MCFS) which performs firstly spectral clustering to get cluster labels and then makes the supervised feature selection with respect to the determined cluster labels.
- Embedded methods include a regularization term to the unsupervised learning problem. In (Wang et al., 2015) authors propose an embedded feature selection framework that incorporates sparse learning in the clustering problem to select features with respect to the cluster labels.

Recently, new unsupervised variable selection methods have emerged, allowing to reduce redundancy in data without label information. These methods are based on the principle of self-representation (Zhu et al., 2015), artificial neural networks (Han et al., 2018), …

In this work, we are interested in problems of unsupervised nonlinear variable selection problems for continuous variables. Using our former work FS-ELM (Kanouté et al., 2023) based on neural networks and proposed to deal with nonlinear feature selection for multi-target regression problems, our contribution is to propose an extension of this approach to the case of unsupervised feature selection problem. Applications to remove noise and redundant variables from the original set of variables on both synthetic and real-world datasets will be introduced to analyze the performances of this new method.

[a] https://orcid.org/0009-0009-6225-2880

Kanouté, M., Grall-Maës, E. and Beauseroy, P.
Unsupervised Feature Selection Using Extreme Learning Machine.
DOI: 10.5220/0013067500003837
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Joint Conference on Computational Intelligence (IJCCI 2024)*, pages 621-628
ISBN: 978-989-758-721-4; ISSN: 2184-3236
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

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 extended in the unsupervised framework is exposed. Experimental results are given and discussed in section 4. Finally, in section 5, conclusions are drawn.

2 NOTATIONS AND RELATED WORKS

2.1 Notations

Considering the problem of unsupervised feature selection, the following notations are used:

- $\mathcal{S}$ is the set of variables.
- X is the matrix of n observations whose variables are in $\mathcal{S}$. It is supposed that X is normalized, that is to say, that for each variable of X, the mean is 0 and the variance is 1.
- 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 $M \in \mathbb{R}^{n \times d}$ (matrix of n rows and d columns), 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 first applies the l_2 norm to each row of the matrix and then applies the l_1 norm to the computed norm. This norm, therefore, makes it possible to impose sparsity on the rows.

- 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:

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

2.2 Related Works

In this section, some unsupervised feature selection methods related to this work are described. In (Zhu et al., 2015) the authors propose a regularized self-representation (RSR) model for unsupervised feature selection. It is based on the principle of self-representation (where each feature can be represented as the linear combination of its relevant features) and $l_{2,1}$-norm regularization. RSR makes it possible to remove redundancy in $\mathcal{S}$ by selecting important features that participate in the representation of most of the other features. The representation coefficients matrix noted $W^{(1)}$ is determined by minimizing the following expression:

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

C is the regularization parameter for sparsity. The larger is C the sparser is $W^{(1)}$. This parameter tunes the trade-off between the reconstruction loss and the number of selected variables. Once C^* the optimal value has been determined according to a criterion, the importance of each variable is determined by calculating the Euclidean norm of its corresponding row in $W^{(1)}$, and variables with low weight can be removed. Only the linear relationships between the variables are exploited by this approach. In (Han et al., 2018), the authors propose AutoEncoder-inspired unsupervised Feature Selection (AEFS), a nonlinear approach based on a single hidden layer auto-encoder and a $l_{2,1}$-norm regularization term on the weight matrix of the hidden layer to select relevant features while reconstructing the network inputs. The expression to be optimized is:

$$\mathcal{L}_C(\Theta) = \frac{1}{2n} ||X - \widehat{X}||_F^2 + C||W^{(1)}||_{2,1} + \frac{\lambda}{2} \sum_{i=1}^{2} ||W^{(i)}||_F^2 \quad (5)$$

where

- n is the number of observations for training.
- $\Theta = \{W^{(1)}, W^{(2)}\}$ 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.
- $\widehat{X} = \sigma(XW^{(1)})W^{(2)}$ where σ is an activation function.
- C is a regularization parameter for sparsity (as defined in Equation 4).
- λ is a regularization parameter allowing stability and promoting convergence.

Once the optimal couple (C^*, λ^*) has been determined according to a criterion, the importance of each variable i is determined by calculating the Euclidean norm of its corresponding row in $W^{(1)}$ i.e. $||W_i^{(1)}||_2$. Although AEFS exploits nonlinear relationships unlike RSR, one of its limitations is due to the simplicity of the model. Indeed, AEFS is composed of a single hidden layer with a number of neurons smaller than the number of input variables, which could not capture the complex nonlinear relationships between

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

features. In (Mirzaei et al., 2020) the authors propose Unsupervised Teacher-Student Feature Selection (U-TSFS) an approach based on knowledge distillation. Two models called teacher and student networks are considered. The teacher model is a complex nonlinear method such as deep auto-encoder or manifold learning techniques (PCA (Hotelling, 1933), TSNE (Van der Maaten and Hinton, 2008), ISOMAP (Tenenbaum et al., 2000)) which tries to obtain the best low dimensional representation of the data denoted $L \in \mathbb{R}^{n\times l}$ with $l << d$ defined as follows:

$$L = F(X) \qquad (6)$$

where F is the complex nonlinear function model such as a deep autoencoder or manifold learning techniques.

The student model is a simple single-layer neural network in which a $l_{2,1}$-norm regularization term is added to the weight matrix of the hidden layer to select relevant features while trying to mimic the low dimensional L. Hence the feature selection is done with a simple hidden layer so that the error can be easily back-propagated and the relevant features selected efficiently. For better training in the student model, the low representation L has been normalized between 0 and 1 as follows $(L_{sc})_{ij} = \frac{L_{ij}-min(L^j)}{max(L^j)-min(L^j)}$ for $1 \leq i \leq n$ and $1 \leq j \leq l$. The expression to be minimized in the student network is:

$$\frac{1}{2n}||L_{sc}-\widehat{L}_{sc}||_F^2 + C||W^{(1)}||_{2,1} \qquad (7)$$

where $\widehat{L}_{sc} = Relu(XW^{(1)}+b^{(1)})W^{(2)}+b^{(2)}$.

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

RSR, NFSN, and U-TSFS are methods that allow unsupervised variable selection. RSR exploits only linear relationships between variables while AEFS and U-TSFS exploit nonlinear relationships between variables. In 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.

3 PROPOSED APPROACH

In this part, the formulation of FS-ELM method proposed to deal with multi-target feature selection problems is introduced first and its extension to the unsupervised case is then developed. In section 4, the proposed extension is assessed.

3.1 Feature Selection Using Extreme Learning Machine (FS-ELM)

Feature Selection Using Extreme Learning Machine (FS-ELM) (Kanouté et al., 2023) is an approach that determines relevant features based on nonlinear multi-output regression. The feature selection is done by training a regression model using Extreme Learning Machine (ELM) (Schmidt et al., 1992) which is a type of neural network with one hidden layer with randomly generated weights $W^{(1)}$, and an output layer in which the weights $W^{(2)}$ are updated. The feature selection idea consists of associating to each feature i a weight $\alpha_i \in [0,1]$ to be tuned during the training of ELM. This model has been proposed first in (Challita et al., 2016) for a two-class classification problem. Figure 1 illustrates the architecture of this model.

Let n be the sample size and p the number of variables.

Let $Nneur$ be the number of neurons in the hidden layer.

Let $X = [a_1, \cdots, a_p] \in \mathbb{R}^{N\times p}$ where $a_i \in \mathbb{R}^N$ is the realisation of feature i for all observations and $Y \in \mathbb{R}^{N\times c}$ a matrix containing the target variables ($c > 1$).

The selection of features is done by minimizing with respect to Θ, the following expression:

$$\mathcal{L}_{\lambda,C}(\Theta) = ||Y - Y_\Theta||_F^2 + \lambda||W^{(2)}||_F^2 + C\sum_{i=1}^{p}\alpha_i \qquad (8)$$

where

- $\Theta = (W^{(2)},\ \alpha = (\alpha_1,\ldots,\alpha_p))$ are the parameters to be optimized.
- $Y_\Theta = S_\alpha W^{(2)} \in \mathbb{R}^{N\times c}$ is the network output.
  - $W^{(2)} \in \mathbb{R}^{Nneur\times c}$ is the weight matrix of the network output also including a bias.
  - $S_\alpha = \sigma[X_\alpha W^{(1)}]$ where
    - σ is an activation function.
    - $W^{(1)} \in \mathbb{R}^{(p+1)\times Nneur}$ is the weight matrix of the hidden layer that includes a bias coefficient. It is a random matrix.
    - $X_\alpha = X^{'}D_\alpha$ is a $N\times(p+1)$ matrix where
      - $X^{'} = \begin{pmatrix} X^T \\ \mathbf{1}_N^T \end{pmatrix}^T$ is $N\times(p+1)$ matrix with $\mathbf{1}_N$ 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 to each variable such that $(D_\alpha)_{i,i} = \alpha_i$ with $\alpha_i \in [0,1]$ the weight associated to each variable i for $i = 1,\cdots,p$ and α_{p+1} is the weight associated to the fixed input (bias) arbitrarily set to 1 i.e. $\alpha_{p+1} = 1$.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

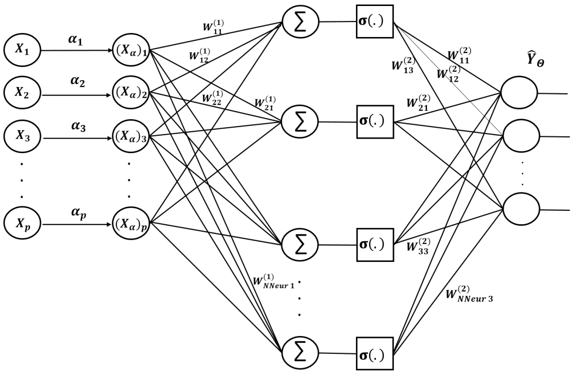


Figure 1: Architecture of the used approach.

- C is the regularization parameter for sparsity that allows setting some α_i to 0.
- λ is the regularization parameter allowing improvement of stability and promoting convergence.

3.2 Determination of Parameters

To select features for estimating the target variables, the optimal values of $\Theta = (W^{(2)}, \alpha)$ are determined using an optimization strategy that consists of updating them alternately and iteratively. That is, $W^{(2)}$ is updated with fixed D_α and vice versa.
For a given value α, $W^{(2)}$ is updated by calculating the derivative of Equation 8 with respect to $W^{(2)}$, which leads to the simple closed form solution:

$$W^{(2)} = (S_\alpha S_\alpha^T + \lambda I)^{-1} S_\alpha^T Y \tag{9}$$

For fixed $W^{(2)}$, $\alpha = (\alpha_1, \dots, \alpha_p)$ is updated such that $\alpha_i \in [0,1]$. In our former work, the partial derivative of Equation 8 with respect to α_i is approximated using numerical methods. The optimization problem can be reformulated as:

$$\begin{aligned} \underset{\alpha_i}{\text{minimize}} \quad & \mathcal{L}_{\lambda,C}(\Theta) \\ \text{subject to} \quad & \alpha_i \in [0,1] \quad \text{for } i = 1, \dots, p. \end{aligned} \tag{10}$$

3.3 Unsupervised Feature Selection Using Extreme Learning Machine (U-FS-ELM)

The effectiveness of FS-ELM for multi-target variable selection has been shown on synthetic and real data. The proposed approach called U-FS-ELM meaning Unsupervised Feature Selection Using Extreme Learning Machine is an extension of our former work in the unsupervised case by taking $Y = X$. Unsupervised feature selection in U-FS-ELM is performed by minimizing with respect to Θ the following expression:

$$\mathcal{L}_{\lambda,C}(\Theta) = ||X - X_\Theta||_F^2 + \lambda ||W^{(2)}||_F^2 + C \sum_{i=1}^{p} \alpha_i \tag{11}$$

Unlike the nonlinear approaches mentioned above (AEFS, U-TSFS) which determine the important features from a low-dimensional representation, this approach has the advantage of being feedforward in addition to associating weights between 0 and 1 to each feature according to their importance in the data (linear or non-linear relationship with other variables). Thus by choosing a very large number of neurons *Nneur*, the input variables can be correctly estimated as stated in the universal approximation theorem (Hornik, 1991) and the addition of sparsity regularisation allows variable selection.

4 EXPERIMENTS

In this part, the original set $\mathcal{S}$ of the features is assumed to be defined. The determination of a subset of relevant variables $\mathcal{S}' \subset \mathcal{S}$ by U-FS-ELM for the reconstruction of the variables of $\mathcal{S}$ is assessed.
The subsections 4.1 and 4.2 concern respectively the evaluation of U-FS-ELM on synthetic data and real-world data. U-FS-ELM with $Nneur = 400$ is compared with the following approaches:

- RSR
- AEFS with $Nneur \in \{\lfloor \frac{p}{2} \rfloor + 1, p-1\}$ where p is the number of variables and $\lfloor \frac{p}{2} \rfloor$ is the floor of $\frac{p}{2}$ i.e. the greatest integer less than or equal to $\frac{p}{2}$.
- U-TSFS where the teacher model F is $TSNE$ with the number of components $n\_comp \in \{2,3\}$ and in the student network $Nneur = n\_comp \times 10$.

To assess the proposed method and compare it with other methods, the relevance of the selected features for each approach is assessed. To avoid the bias problem during the assessment, the original dataset D has been split into two subsets D_{train} (67% of D) and D_{test} (33% of D). Optimal values of hyperparameters have been determined according to a criterion on the MSE by 5-fold cross-validation on D_{train} as follows:

- for $C \in I_C = \{10^{-4}, 10^{-3}, \dots, 10^3, 10^4\}$ and for $\lambda \in I_\lambda = \{10^{-4}, 10^{-3}, \dots, 10^3, 10^4\}$.
Compute $\widehat{Y}^{(\lambda,C)}$ the estimate of Y associated with C and λ and $MSE(Y, \widehat{Y}^{(\lambda,C)})$.
- Choose $(C^*, \lambda^*) \in I_C \times I_\lambda$ such that

$$(C^*, \lambda^*) = \underset{(C,\lambda) \in I_C \times I_\lambda}{\text{argmin}} \; MSE(Y, \widehat{Y}^{(\lambda,C)}) < 0.1 \tag{12}$$

SCIENCE AND TECHNOLOGY PUBLICATIONS

obtained by 5 fold cross-validation on D_{train}.
For RSR and U-TSFS the procedure is similar to the one above but only C^* is determined.

Once the optimal hyperparameters have been chosen, the feature weights have been determined by running the feature selection approach on all observations of D_{train} using the optimal hyperparameters. Then the importance of each feature has been determined using its representation in the feature weights as follows:

- for RSR, AEFS, and U-TSFS, rank the variables as defined in section 2.2.
- for U-FS-ELM, rank the variables according to the scaling factors α_i.

Once features have been ranked, the pertinence has been assessed by building p models on D_{train} and evaluating them on D_{test} by keeping from 1 to p variables corresponding to the highest rank for reconstructing all variables of $\mathcal{S}$. The model used for evaluation is a one-hidden-layer neural network with 500 neurons. The activation function is sigmoid and the optimizer is adam. The metric used for the assessment of the model is the MSE.
The observations of all variables have been normalized (removing the mean and scaling to unit variance) to avoid scaling problems.

4.1 Synthetic Dataset

This section describes the results with two generated datasets called **synth1** and **synth2**. In **synth1**, there are only linear relationships (coefficients have been randomly determined according to a continuous uniform distribution between -1 and 1) between features while in **synth2** there are nonlinear relationships between features. In **synth1** (resp. **synth2**), 7 (resp. 8) features were firstly defined then 5 (resp. 7) random features that depend on these 7 (resp. 8) features with linear (resp. nonlinear) relationships were defined. Finally, in **synth1**, 3 redundant features were created from the 7 features first defined. Thus 2000 observations have been generated from these 15 features for each dataset. In **synth1**, the variables are defined as follows :
$f_1, f_2, \ldots, f_7 \sim \mathcal{N}(0,1)$
$f_8 = -0.56f_1 + 0.22f_2 - 0.84f_3 - 0.46f_4 + 0.2f_5 - 0.72f_6 - 0.96f_7 + \varepsilon_8$
$f_9 = 0.74f_1 + 0.54f_2 + 0.48f_3 - 0.18f_4 - 0.46f_5 - 0.66f_6 - 0.6f_7 + \varepsilon_9$
$f_{10} = -0.58f_1 + 0.04f_2 - 0.12f_3 - 0.4f_4 - 0.44f_5 + 0.92f_6 + 0.4f_7 + \varepsilon_{10}$
$f_{11} = 0.84f_1 - 0.4f_2 - 0.68f_3 + 0.26f_4 - 0.5f_5 + 0.92f_6 + 0.56f_7 + \varepsilon_{11}$
$f_{12} = -0.02f_1 - 0.62f_2 + 0.76f_3 + 0.16f_4 - 0.34f_5 - 0.62f_6 - 0.96f_7 + \varepsilon_{12}$
$f_{13} = f_7 + \varepsilon_{13}$, $f_{14} = f_3 + \varepsilon_{14}$, $f_{15} = f_5 + \varepsilon_{15}$. $\varepsilon_8, \varepsilon_9, \varepsilon_{10}, \varepsilon_{11}\varepsilon_{12} \sim \mathcal{N}(0, 0.05)$ and ε_{13}, ε_{14}, $\varepsilon_{15} \sim \mathcal{N}(0,1)$.

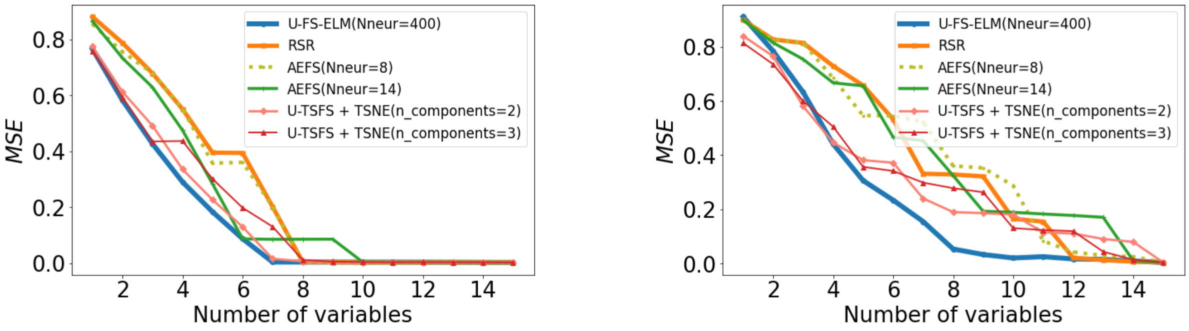
For **synth2**, the variables are defined as follows:
$f_1 \sim \mathcal{N}(1,0.5)$; $f_2 \sim \mathcal{N}(0.7,1)$; $f_3 \sim \mathcal{N}(3,1)$; $f_4 \sim \mathcal{N}(0,0.5)$; $f_5 \sim \mathcal{N}(0.3,1)$; $f_6 \sim \mathcal{N}(2,0.7)$; $f_7 \sim \mathcal{U}(-1,1)$; $f_8 \sim \mathcal{U}(-3,1)$
$f_9 = f_1 sin(f_1) + \varepsilon_{f_9}$ where $\varepsilon_{f_{10}} \sim \mathcal{N}(0,0.08)$
$f_{10} = f_2^3 + 2f_2 + e^{f_2 - f_4 - f_6^2} - cos(f_4 - f_6 + f_2) + \varepsilon_{10}$ where $\varepsilon_{f_{10}} \sim \mathcal{N}(0,0.08)$
$f_{11} = e^{f_1 - f_4^2} + \varepsilon_{f_{11}}$ where $\varepsilon_{f_{11}} \sim \mathcal{N}(0,0.1)$
$f_{12} = \dfrac{|f_3 + f_4|}{f_3^2 + f_4^2} + \varepsilon_{f_{12}}$ where $\varepsilon_{f_{12}} \sim \mathcal{N}(0,0.04)$
$f_{13} = e^{f_4} cos(f_6) + \varepsilon_{f_{13}} \sim \mathcal{N}(0,0.02)$
$f_{14} = arctanh(f_7) + \varepsilon_{f_{14}}$ where $\varepsilon_{f_{14}} \sim \mathcal{N}(0,0.08)$
$f_{15} = ln(3 - f_8^2 - 2f_8) + arctan(\sqrt{3 - f_8}) + \varepsilon_{f_{15}}$ where $\varepsilon_{f_{15}} \sim \mathcal{N}(0,0.08)$

For the two datasets, the goal was to determine the subset $\mathcal{S}^{'} \subset \mathcal{S}$ such that the reconstruction loss between the variables of $\mathcal{S}$ and their estimate from the variables of $\mathcal{S}^{'}$ is minimized. The optimal values for the hyperparameters have been determined by 5-fold cross-validation for each approach as described above. The chosen parameters on each dataset for each approach are given in Table 1. After the choice of regularization parameters on **synth1** and **synth2**, variables have been ranked according to their importance for each approach, and the list is given in Table 2. It may be noticed that U-FS-ELM manages to better select independent features compared to RSR, AEFS, and U-TSFS, in particular on **synth1** it set all coefficients to zero after the 7th most important variable. Figure 2 shows the estimated value of the mean of MSE between X and its estimated value $\widehat{X}$ versus the number of most important variables used to build the model.

Table 1: Chosen parameters on each synthetic dataset for each approach.

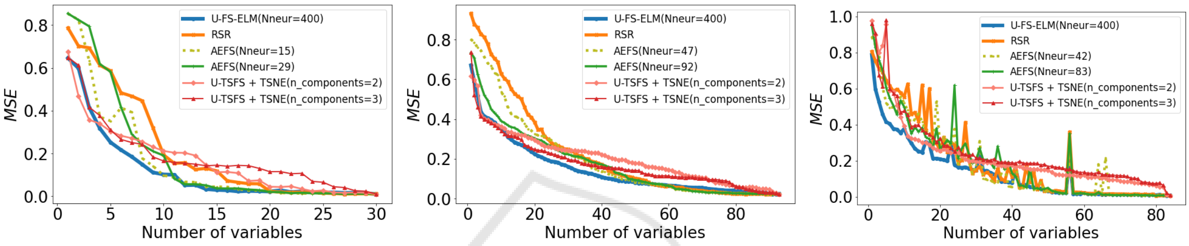
	synth1		**synth2**	
Methods	λ	C	λ	C
U-FS-ELM (Nneur = 400)	1	10^3	10^{-1}	10^3
RSR	-	10^2	-	10^2
AEFS (Nneur = 8)	10^{-3}	10^{-2}	10^{-4}	10^{-2}
AEFS (Nneur = 14)	10^{-4}	10^{-1}	10^{-3}	10^{-2}
U-TSFS (n_comp = 2)	-	10^{-3}	-	10^{-3}
U-TSFS (n_comp = 3)	-	10^{-3}	-	10^{-3}

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS



(a) **synth1** (b) **synth2**

Figure 2: MSE versus the number of most important variables on synthetic data.



(a) Breast cancer data set (b) Ansur data II (c) US crimes dataset

Figure 3: MSE versus the number of most important variables on real-world datasets.

Table 2: List of ranked variables for each approach.

(a) **synth1** dataset

				AEFS				U-TSFS+TSNE			
U-FS-ELM		RSR		Nneur = 8		Nneur = 14		n_comp = 2		n_comp = 3	
var	weight	var	weight	var	weight	var	weight	var	weight	var	weight
f_7	$5,5.10^{-2}$	f_4	$7,5.10^{-1}$	f_1	$2,5.10^{-1}$	f_2	$4,5.10^{-4}$	f_3	$4,9.10^{-1}$	f_{12}	$2,5.10^{-1}$
f_3	$5,3.10^{-2}$	f_2	$6,3.10^{-1}$	f_2	$1,9.10^{-1}$	f_6	$2,9.10^{-4}$	f_{10}	$3,4.10^{-1}$	f_{14}	2.10^{-1}
f_5	5.10^{-2}	f_1	$3,9.10^{-1}$	f_4	$1,6.10^{-1}$	f_1	$2,4.10^{-4}$	f_1	$2,6.10^{-1}$	f_{11}	$1,6.10^{-1}$
f_6	$4,8.10^{-2}$	f_6	$2,9.10^{-1}$	f_6	$1,6.10^{-1}$	f_5	10^{-4}	f_{13}	$2,4.10^{-1}$	f_3	$1,3.10^{-1}$
f_1	$4,7.10^{-2}$	f_{15}	$1,6.10^{-1}$	f_3	6.10^{-2}	f_3	$8,9.10^{-5}$	f_{11}	$2,3.10^{-1}$	f_5	$1,3.10^{-1}$
f_2	$4,6.10^{-2}$	f_5	$1,4.10^{-1}$	f_{14}	4.10^{-2}	f_7	$6,9.10^{-5}$	f_8	$2,2.10^{-1}$	f_{13}	$1,2.10^{-1}$
f_4	$3,9.10^{-2}$	f_{14}	$1,3.10^{-1}$	f_5	$3,2.10^{-2}$	f_{15}	$5,3.10^{-5}$	f_5	$1,4.10^{-1}$	f_8	$1,1.10^{-1}$
f_8	0	f_{13}	$1,2.10^{-1}$	f_7	3.10^{-2}	f_{14}	$3,8.10^{-5}$	f_4	$1,2.10^{-1}$	f_1	$9,2.10^{-2}$
f_9	0	f_3	$9,4.10^{-2}$	f_{15}	3.10^{-2}	f_{13}	$2,2.10^{-5}$	f_2	$1,1.10^{-1}$	f_2	$8,4.10^{-2}$
f_{10}	0	f_7	$7,8.10^{-2}$	f_{13}	$1,8.10^{-2}$	f_{10}	$2,2.10^{-5}$	f_{14}	$1,1.10^{-1}$	f_{10}	$7,3.10^{-2}$
f_{11}	0	f_8	$8,8.10^{-3}$	f_9	$2,1.10^{-3}$	f_8	$1,9.10^{-5}$	f_{15}	10^{-1}	f_7	$6,4.10^{-2}$
f_{12}	0	f_{10}	$7,6.10^{-3}$	f_8	$1,6.10^{-3}$	f_{11}	$8,4.10^{-6}$	f_{12}	$8,2.10^{-2}$	f_9	$5,8.10^{-2}$
f_{13}	0	f_9	$6,8.10^{-3}$	f_{10}	$1,4.10^{-3}$	f_{12}	$1,5.10^{-6}$	f_6	$6,3.10^{-2}$	f_6	$5,8.10^{-2}$
f_{14}	0	f_{11}	$5,8.10^{-3}$	f_{11}	$9,1.10^{-4}$	f_9	$1,5.10^{-6}$	f_7	$2,1.10^{-2}$	f_{15}	$3,8.10^{-2}$
f_{15}	0	f_{12}	$4,1.10^{-3}$	f_{12}	$5,4.10^{-4}$	f_4	6.10^{-8}	f_9	$1,9.10^{-2}$	f_4	$2,9.10^{-2}$

(b) **synth2** dataset

				AEFS				U-TSFS + TSNE			
U-FS-ELM		RSR		Nneur = 8		Nneur = 14		n_comp = 2		n_comp = 3	
var	weight	var	weight	var	weight	var	weight	var	weight	var	weight
f_4	$1,2.10^{-1}$	f_8	1	f_8	$2,5.10^{-2}$	f_8	$3,1.10^{-1}$	f_{10}	$4,9.10^{-1}$	f_{11}	$2,4.10^{-1}$
f_6	$7,9.10^{-2}$	f_5	1	f_5	$2,3.10^{-2}$	f_4	2.10^{-1}	f_{12}	$2,9.10^{-1}$	f_{12}	$1,8.10^{-1}$
f_{10}	$6,6.10^{-2}$	f_{15}	1	f_{15}	$1,4.10^{-2}$	f_5	$1,7.10^{-1}$	f_{11}	$2,2.10^{-1}$	f_{13}	$1,4.10^{-1}$
f_1	$2,5.10^{-2}$	f_4	1	f_{14}	4.10^{-3}	f_3	$1,4.10^{-1}$	f_{13}	$1,3.10^{-1}$	f_8	$1,3.10^{-1}$
f_{14}	$2,2.10^{-2}$	f_3	1	f_{13}	$2,7.10^{-3}$	f_{15}	$1,1.10^{-1}$	f_4	$1,2.10^{-1}$	f_2	$1,3.10^{-1}$
f_5	$1,8.10^{-2}$	f_6	1	f_7	$2,6.10^{-3}$	f_9	$3,8.10^{-2}$	f_1	$1,1.10^{-1}$	f_{15}	$1,1.10^{-1}$
f_3	$1,8.10^{-2}$	f_{11}	1	f_6	$2,4.10^{-3}$	f_{11}	$3,4.10^{-2}$	f_{14}	$8,8.10^{-2}$	f_3	$9,9.10^{-2}$
f_8	$1,8.10^{-2}$	f_{13}	1	f_{10}	$2,2.10^{-3}$	f_6	$2,9.10^{-2}$	f_{15}	$8,7.10^{-2}$	f_1	$8,8.10^{-2}$
f_{12}	$1,8.10^{-2}$	f_{12}	1	f_2	$1,8.10^{-3}$	f_{14}	$2,9.10^{-2}$	f_9	$8,5.10^{-2}$	f_{10}	$8,8.10^{-2}$
f_{15}	$1,7.10^{-2}$	f_2	1	f_{12}	$1,7.10^{-3}$	f_1	$2,6.10^{-2}$	f_7	7.10^{-2}	f_{14}	8.10^{-2}
f_2	0	f_{10}	1	f_1	$1,3.10^{-3}$	f_{12}	$2,5.10^{-2}$	f_8	$4,6.10^{-2}$	f_7	$6,6.10^{-2}$
f_7	0	f_7	1	f_3	$9,7.10^{-4}$	f_{13}	$2,1.10^{-2}$	f_2	$3,4.10^{-2}$	f_9	$2,6.10^{-2}$
f_9	0	f_{14}	1	f_{11}	$6,9.10^{-4}$	f_7	2.10^{-2}	f_3	3.10^{-2}	f_5	$2,5.10^{-2}$
f_{11}	0	f_9	1	f_9	$5,9.10^{-4}$	f_2	$1,6.10^{-2}$	f_6	$2,1.10^{-2}$	f_4	$2,1.10^{-2}$
f_{13}	0	f_1	1	f_4	$2,2.10^{-4}$	f_{10}	$1,5.10^{-2}$	f_5	$1,5.10^{-3}$	f_6	$1,4.10^{-2}$

4.2 Real-World Datasets

This part presents the results on real-world datasets. Table 3 contains the list of real-world data used as well as the number of variables, and the number of samples. Some information about these datasets as well as the pre-processing done are described below:

Table 3: Real-world data sets.

Name	Size	Features	Source
Breast Cancer	569	30	(Zwitter and Soklic, 1988)
Ansur data 2	6068	93	(Paquette et al., 2009)
US crimes	2215	125	(Redmond, 2009)
Mnist	10000	784 (28×28)	(Deng, 2012)

Table 4: Number of variables with weights greater than 0 and percentage of selected variables among the 784 variables for different values of C with $\lambda = 10^{-2}$.

C	Number of variables with $\alpha_i > 0$	% of selected features
10^{-2}	457	58.29 %
10^{-1}	365	46.56 %
1	208	26.53 %
10^{1}	106	13.52 %

- Breast cancer

The breast cancer dataset contains 569 observations of features extracted from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. The goal is to determine if breast cancer is cancerous or non-cancerous based on extracted features. This dataset is often used to explore feature selection techniques. In this paper, the nominal variable containing the classes is removed. U-FS-ELM and other unsupervised feature selection tech-

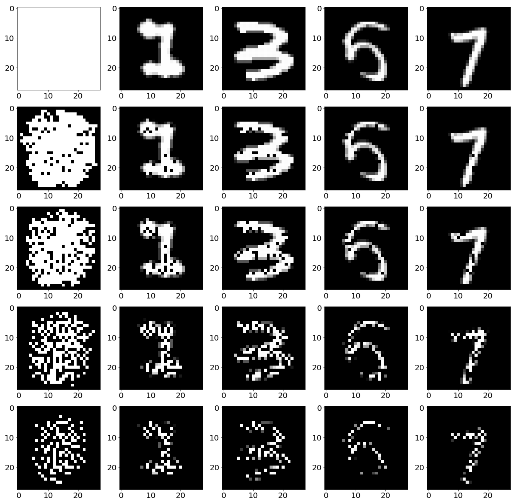


Figure 4: Mnist dataset image reconstruction using 200 images as the training dataset. The first row contains the retained pixels (first column) and some original images (second column to last column). The following 4 rows contain respectively the retained pixels by U-FS-ELM (first column) and the reconstruction results (second column to last column) using the retained pixels for $C = 10^{-2}$,$C = 10^{-1}$,$C = 1$,$C = 10^{1}$ with $\lambda = 10^{-2}$.

niques have been applied to the normalized observations of 29 variables to select important features and remove the redundant variables. The dataset can be downloaded on the UCI website at this URL: https://archive.ics.uci.edu/datasets.

- Ansur data II

The Anthropometric Survey of US Army Personnel (ANSUR 2 or ANSUR II) is a database with measurements of American military personnel done in 2012 and made public in 2017. This database contains 6000 observations (4082 men and 1986 women) of 93 numerical anthropometric measurements that describe the size and shape of the human. The feature selection methods have been applied to the 1145 observations of females whose age is between 20 and 30. The dataset is available at this URL: https://www.openlab.psu.edu/ansur2/.

- Communities and Crime

It is a dataset on crime in communities in the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data, and crime data with a target variable and 127 other variables. Among the 127 remaining variables, 5 variables are considered as non-predictive in the description of the dataset. After removing the variables with missing values and non-predictive variables, the feature selection approaches have been applied to 84 continuous variables to select important features allowing the reconstruction of all of them. The data set is available on the UCI website.

- Mnist (Mixed National Institute of Standards and Technology)

This dataset is composed of 10000 black and white handwritten digits images used for training neural network in computer visions. It contains ten classes corresponding to the 10 numerical digits. For each handwritten digit, there are 28 × 28 pixels between 0 and 255. To apply our approach, the dataset has been normalized by the min-max feature scaling method which brings all values between 0 and 1 and is defined for each variable as $x^{'} = \dfrac{x - min(x)}{max(x) - min(x)}$ where x contains observations of a variable of $\mathcal{S}$. To determine the feature weights, only 200 observations have been considered as training dataset and another 200 observations as the test dataset.

After the choice of the regularization parameters and the ranking of the features for each approach, X has been reconstructed by building p models from 1 to p variables corresponding to the highest rank. The number of important variables taken successively is $\{1,2,\ldots,30\}$ on Breast cancer data set, $\{1,2,\ldots,93\}$ on Ansur data II set, $\{1,2,\ldots,84\}$ on Communities and Crime data set. Figure 3 shows the MSE between X and its estimated value $\widehat{X}$ versus the number of important variables taken successively on these datasets and it can be noticed that U-FS-ELM performs well compared to AEFS and U-TSFS, precisely on Breast cancer dataset between 1 and 10 first important variables, on Ansur data II dataset between the first 15 and the first 40 important variables, and on Communities and Crime dataset U-FS-ELM has the lowest MSE for any number of variables taken.

The proposed method successfully reduces the initial number of variables in structured continuous data by keeping relevant variables that can estimate properly other related variables. It can also be noticed that generally, U-FS-ELM selects important variables better than AEFS and U-TSFS. Indeed, if the number of variables p is not very large, AEFS is a simple autoencoder that may not capture complex relationships between features, and in the U-TSFS approach the teacher model must be chosen according to the data to obtain a better representation of the data in order to avoid propagating the estimation errors in the student model and this latter also requires the choice of the right activation function, the number of neurons, the choice of the optimal parameter $C^{*}, \ldots$

U-FS-ELM has been also applied to image data, the mnist dataset. The goal was to determine relevant variables among 784 (28 × 28) variables corresponding to pixels on each handwritten digit. U-FS-ELM with $\lambda = 10^{-2}$ and $C \in \{10^{-2}, 10^{-1}, 1, 10^{1}\}$ was trained on 200 images (20 images per class) randomly chosen and the number of variables with weights greater than 0 and percentage of selected variables among the 784 variables for each value of C is given in Table 4.

The reconstruction results using these hyperparameter values for some images in the test dataset are shown in Figure 4 and it can be noticed that U-FS-ELM has reduced the number of features while keeping useful information. It should be noted that this approach is different from reduction methods which determine a representation of the data in a subspace while here a selection of important variables is done.

5 CONCLUSIONS

In this paper, an approach is proposed to deal with unsupervised feature selection problems exploiting nonlinear relationships between variables. It consists of assigning to each feature i a weight $\alpha_i \in [0,1]$ updated during the reconstruction of the input variables and of determining hyperparameters λ and C which are respectively parameters for stability and sparsity. By tuning these hyperparameters according to MSE, the weights α_i associated to the features make it possible to determine important features while minimizing the reconstruction error. Many experiments have been done on two synthetic data, three structured continuous real-world data, and one image data and the results have been compared with other methods. They show the effectiveness of the proposed approach.

REFERENCES

Cai, D., Zhang, C., and He, X. (2010). Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342.

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

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

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*.

Han, K., Wang, Y., Zhang, C., Li, C., and Xu, C. (2018). Autoencoder inspired unsupervised feature selection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2941–2945. IEEE.

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

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Kanouté, M., Grall-Maës, E., and Beauseroy, P. (2023). Neural network-based approach for supervised nonlinear feature selection. In *Proceedings of the 15th International Joint Conference on Computational Intelligence - Volume 1: NCTA*, pages 431–439. INSTICC, SciTePress.

Mirzaei, A., Pourahmadi, V., Soltani, M., and Sheikhzadeh, H. (2020). Deep feature selection using a teacher-student network. *Neurocomputing*, 383:396–408.

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

Paquette, S., Gordon, C. C., and Bradtmiller, B. (2009). Anthropometric survey (ansur) ii pilot study: Methods and summary statistics.

Redmond, M. (2009). Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

Schmidt, W., Kraaijveld, M., and Duin, R. (1992). Feedforward neural networks with random weights. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, pages 1–4.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wang, S., Tang, J., and Liu, H. (2015). Embedded unsupervised feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Zhu, P., Zuo, W., Zhang, L., Hu, Q., and Shiu, S. C. (2015). Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446.

Zwitter, M. and Soklic, M. (1988). Breast Cancer. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51P4M.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS