Feature selection in Solar Radiation prediction using Bootstrapped SVRs

O. García-Hinde[†], V. Gómez-Verdejo[†], M. Martínez-Ramón[†‡],
C. Casanova-Mateo[a*], J. Sanz-Justo[*], S. Jiménez-Fernández[§] and S. Salcedo-Sanz [§]

[†]Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Spain.
[‡]Department of Electrical and Computing Engineering, The University of New Mexico, NM, USA.
[a] Department of Civil Engineering: Construction, Infrastructure and Transport, UPM, Madrid.
[*]LATUV, Universidad de Valladolid, Valladolid, Spain.
[§]Department of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain.
Contact e-mail: oghinde@tsc.uc3m.es

***Abstract*—During the past years solar radiation prediction has become increasingly relevant among the scientific community and Machine Learning techniques have proven to be a useful tool to automatically learn an accurate prediction model. In this paper, we move one step further and try to gain interpretability during the learning process by introducing a novel feature selection approach. Our method trains a set of bootstrapped SVR classifiers to detect those features that are informative for the prediction task. This way we obtain a more robust set of selected features compared to other selection methods. This allows us to detect in a multivariate fashion not only the features needed to solve the prediction task, but also those that are informative for the problem at hand. The application of this algorithm to a Weather Research and Forecasting model, and its comparison to some state of the art tools, shows the advantages of the proposed method both in terms of resistance to overfitting, selection consistency and interpretability, while at the same time improving performance in terms of prediction accuracy.**

I. Introduction

Solar energy is experiencing a growing demand all over the World. According to the Solar Industry Energies Association (SEIA), it is expected that up to 20,000 MW of solar capacity will come online over 2016 and 2017, which will add to the 20,000 MW of solar capacity already installed in the United States[1]. Installed world wide, there are 180,000 MW of solar energy according to the association Solar Power Europe[2].

Solar energy has a highly stochastic behavior caused by the atmospheric conditions around the power facilities. Therefore an optimal usage of this energy requires a high level of forecasting accuracy. The amount of solar power generated by a power plant is a function of the solar radiation. The stochastic component of this variable depends on the cloud coverage and characteristics as well as other factors such as the presence of light absorbing particles in the air [1].

For this reason, the most widely used approach to address the task of solar radiation prediction consists in the physical modeling of the deterministic part of solar radiation by computing the relative position of the sun with respect to the facility in order to obtain a clear sky model, and then adding the atmospheric conditions, including rain, wind speed and other variables [2].

A step ahead in prediction consists of the use of numerical weather models. In [3], [4] the authors applied the Weather Research and Forecasting (WRF) meso-scale method [5] in wind speed prediction, which provides a prediction of atmospherical variables at 96 different heights in a given zone, that can be used as inputs in a prediction system. The present work makes use of this approach.

There is a significant amount of work devoted to the prediction of solar radiation. Most approaches tackle the problem from a computational intelligence perspective. Indeed, machine learning (ML) models have been used for example in [6], where a Radial Basis Function was used in solar radiation prediction in a power plant using weather data. In [7], a Multilayer Perceptron (MLP) is used for daily solar radiation prediction from different combinations of weather data. In [8], a Multilayer Perceptron combined with a genetic algorithm is used to perform solar radiation forecast, but in this work authors use satellite images to perform radiation prediction in large areas in Spain. A comparison of prediction techniques can be found in [9] where authors take a time series prediction approach where the input data consists of historical solar radiation data. There, MLPs, Markov chains, Bayesian inference and ARIMA models are compared. Support Vector Machines (SVM) [10] have also been widely used in solar radiation forecast. For example in [11] SVMs are used to predict monthly solar radiation from meteorological data, and the same authors use them in [12] to estimate solar radiation from air temperature. Extreme learning machines (ELM) have also applied to solar prediction using meteorological variables in [13] and in [14], a Kernel Extreme Learning Machine (KELM) has been compared to a kernel SVM. Other works introduce neuro-fuzzy approaches [15], or hybrid models combining ARMA and artificial neural networks [16] to cite some.

The WRF model used in this paper produces a quantity of variables of the order of 10^4, challenging the use of computational intelligence methods since the number of features may

[1]Source: www.seia.org
[2]Source: www.solarpowereurope.org

exceed the number of available datapoints. If a small dataset (in comparison with the number of features) is used, a ML method must include a very strong regularization strategy or it will be guaranteed to overfit. *Feature selection* methods that are able to discriminate the features useful for the prediction task at hand may dramatically improve the performance of the predictor. In this paper we introduce a novel feature selection method that uses a bootstrapping of SVM classifiers to determine the significance of each feature. The method is able to produce an accurate prediction with a number of features about two magnitude orders below the original number of components. Furthermore, the methodology is able to provide an interpretation of the selection that exceeds the ability of standard methods, as can be seen in the results.

As for the structure of this paper, section II provides a review of the current state of the art in feature selection methods. Section III describes the proposed bootstrapped SVM feature selection method in detail. Section IV formulates the solar radiance problem that is tackled by this work. Section V presents the experimental setup and results. Section VI contains a final discussion on the conclusions and future ideas in light of the experimental results.

II. Review of feature selection methods

Datasets often present a very large number of features yet many of these might not be relevant to the learning task. Some might even be noisy and have a negative impact on the performance of the ML model. Therefore it is vital to take advantage of feature selection algorithms that can identify the features that are relevant and necessary for the learning problem. Furthermore, a powerful feature selection algorithm can also provide insight into the informativeness of a given feature or set of features. Following the categorization presented in [17], feature selection methods are ususally grouped into three categories:

- **Filters:** these methods use relevance measurements to analyse how useful each feature is. The selection of features is thus independent from the learning task and happens as a preprocessing stage prior to the training stage. The filter can be used either to produce a variable ranking in terms of the relevance measurement or be combined with algorithms like forward or backward search to obtain subsets of variables. The filtering measurements can also be non-supervised, like the variance of a feature across the training samples, or supervised, like the correlation or the mutual information between a feature and the target output. Filter methods can be very simple and cost-effective in a computational sense, although they can fail to exploit more complex and profound relationships between the features and the learning task.
- **Wrappers:** wrapper feature selectors use ML methods to select relevant feature subsets. They train a ML model with different feature subsets and produce a ranking based on the accuracy obtained with each one. The fact that they work alongside the learning algorithm allows them to obtain feedback from the ML output. When looking for optimal feature subsets exhaustive search methods can be used with small data sets but the problem can quickly become computationally intractable and prone to producing overfitting. This can again be alleviated by the use of forward or backward search algorithms, as well as other methods. An example of a wrapper is the recursive feature elimination model (RFE) presented in [18]. This algorithm applies a backward search algorithm to iteratively eliminate the variables that have the least influence on the SMV margin. The rationale is that these features contribute very little to improving performance.
- **Embedded methods:** Embedded methods are algorithms that are directly integrated into the learning task. Feature selection is applied alongside the ML method, selecting those features that seem to improve performance. Since they are embedded within the learner, their nature depends on the specific classification method used. For instance, in the case of the Lasso method described in [19], a least squared error model paired with an l_1 norm regularisation term will produce a sparse set of selected features by driving the weights of irrelevant features to zero. These methods tend to fall in between filters and wrappers in terms of computational intensity and can also be less prone to overfitting than wrappers.

It is worthy to note that while filters can be applied as univariant or multivariant methods depending on the relevance measurement and search strategy, wrappers and embedded methods are inherently multivariant due to the fact that they evaluate the relevance of subsets of features and their combined effect on the regression task. In this sense, a filter that ranks each single feature in terms of its relevance across the training set is univariant. On the other hand, a filter that measures the relationship between a set of features and the output is a multivariant method.

Another important aspect of feature selection is how redundant features are treated. Wrappers and embedded methods tend to eliminate a great deal of redundant features, keeping only those that are truly needed for the ML task. In other words, these methods tend to search for the smallest feature subset that can offer the best performance. In many cases this is desirable since it greatly simplifies the problem. However, oftentimes redundant features that have an insignificant effect on overall performance can be very informative in terms of problem interpretability. If there are strong relationships between variables of a given set, many wrappers and embedded methods will ignore most of them keeping only those that present the highest degree of usefulness for the ML model. However, for the sake of interpretation, identifying these groups of highly related features can be very useful for the data scientist. In this sense, while sparsity is an often desired quality for a feature selection algorithm, it can sometimes eliminate information that can be very valuable for the human analyst. Also, exploiting these redundant features can help with overfitting since they produce a more consistent and robust model, whereas sparsity can hurt the generalization

capabilities of the method.

Following this idea, in the next section we propose an efficient multivariate feature selection algorithm that is able to provide a ranking where all relevant features are included.

III. Feature selection through Bootstrapped SVRs: a novel approach

In this article we present a feature selection algorithm inspired by the method described in [20], which has already been adapted to a classification scenario in [21] Here, the authors presented a feature selection method that employs an ensemble of SVM classifiers regularised with the l_1 norm. As has been stated, the l_1 norm has the property of imposing sparsity on the input feature space by driving the weights of non-relevant features to zero. For this reason they coined the algorithm as VS-SSVM for Variable Selection via Sparse SVMs. It is obvious that the spirit behind this approach is very similar to that of the Lasso. The motivation for this model is that in many cases, the sparsity acquired using the l_1 norm in the regularization term can prove to be unstable in the face of small variations in the training and validation sets. To alleviate this instability, the method employs an ensemble of sparse SVMs to asses the consistency of the variable selection.

In the case of the algorithm presented in this article, we move away from the sparsity provided by the l_1 norm. This is due to the fact that our goal is to retain the redundant variables in the model for the sake of interpretability (see section II). To accomplish this, we instead use the l_2 norm and analyze by means of a hypothesis test the probability that the weight assigned to each feature by the SVMs in the ensemble is zero.

Since the task tackled in this article is a regression problem, we will specifically use the Support Vector Regression Machine (SVR). Let's consider a training dataset $\langle \mathbf{x}_l, y_l \rangle_{l=1}^{L_{tr}}$, where $\mathbf{x}_l \in \mathbb{R}^D$ are the input samples, $y_l \in \mathbb{R}$ are their associated targets, D is the dimension of the input feature space and L_{tr} is the total number of training samples. The l_2 regularised SVR [22] is a linear regressor whose goal is to find the parameters $\mathbf{w}$ and b which define the function:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{1}$$

such that, for each training sample, it presents a deviation of at most ϵ from its associated target. For this purpose, the SVR solves the following optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{l=1}^{L_{tr}}(\xi_l + \xi_l^*) \\ \text{subject to:} \quad & y_l - (\mathbf{w}^T\mathbf{x}_l + b) \leq \epsilon + \xi_l, \\ & \mathbf{w}^T\mathbf{x}_l + b - y_l \leq \epsilon + \xi_l^*, \qquad \forall l \\ & \xi_l, \xi_l^* \geq 0, \end{aligned} \tag{2}$$

where slack variables ξ_l and ξ_l^* allow the data to present deviations larger than ϵ and the constant C allows control of the trade-off between the flatness of $f(\mathbf{x})$ and the number of allowed errors.

Our algorithm exploits three basic theoretical properties of linear regression with l_2 norm regularisation:

- Isolated features that are *relevant* will have large weights assigned to them.
- If instead of an isolated relevant feature, we have a set of relevant features which are *redundant* among each other, the importance of the weight will be spread out over the weights of all the features in the set. This way, even though each individual weight might be small, the accumulated weight for the set will be large.
- Conversely, features that are *irrelevant* will be assigned a weight of zero. In practice, due to the idiosyncrasies of training data and the presence of noise, irrelevant features will present a very small albeit non-zero weight.

This means that if one was to simply prune those features that present a small weight there exists a risk of eliminating features that are relevant but strongly related to others, and that therefore present a small, spread out weight.

To tackle this problem, we propose the usage of a bootstrapping technique [23] to approximate the sample distribution of the weights. This is accomplished by training a large number of SVRs, each of them using different subsets sampled at random and with replacement from the training data. Thus, by training K SVRs we will obtain a set of K corresponding weight vectors, $\langle \mathbf{w}^{(k)} \rangle_{k=1}^{K}$. Figure 1 shows an example of the weight distributions for two different features, one strongly relevant and the other markedly irrelevant, after bootstrapping 200 SVRs.

We now need to determine which features are irrelevant, i.e. which of the feature-weights are most likely going to be consistently assigned a value of zero. To do so we use a one-sample location test by means of a Student's t-test [24] in which the null hypothesis is that a feature-weight's sample distribution presents a mean of zero. Applying this test to each of the components of the weight space, the p-value of the test will indicate the probability that the weight distribution has a zero mean (lower p-values indicate that the null hypothesis must be rejected and that the associated weight has a non-zero mean).

Since this method uses a bootstrapping of SVRs for feature selection, we will call it the Bootstrapped SVR for Feature Selector algorithm (BS-FS). Algorithm 1 shows the pseudocode for the BS-FS method.

The optimal feature set can be determined in two different ways. One can preset a specific level of confidence and reject those features whose corresponding p-values do not satisfy this criterion. Although it will be very fast, this method has the disadvantage that the confidence level is arbitrarily set and could yield excessively conservative selections, especially when one is working with a large number of features. A different approach is to determine the optimal set of features through cross validation. This approach, while computationally less efficient, will yield a much more significant result towards

Algorithm 1: The BS-FS algorithm.

Data:
$\mathbf{X}_{tr}, \mathbf{y}_{tr}$
Input:
$K =$ number of bootstrap iterations
$S =$ number of samples per subset
for $k \in K$ **do**
  $\mathbf{X}_{tr}^{(k)}, \mathbf{y}_{tr}^{(k)} =$ random sample of size S from $\mathbf{X}_{tr}$ & $\mathbf{y}_{tr}$
  Train SVR using $\mathbf{X}_{tr}^{(k)} \rightarrow$ obtain $\mathbf{w}^{(k)}$
  $\mathbf{W}(k,:) := \mathbf{w}^{(k)}$
for $d \in D$ **do**
  Perform Student's t-test for $\mathbf{W}(:,d)$ with null hypothesis: $\bar{w}_d = 0$
  Store p-value for feature d
Rank features according to their corresponding p-values.

the final performance of the algorithm.

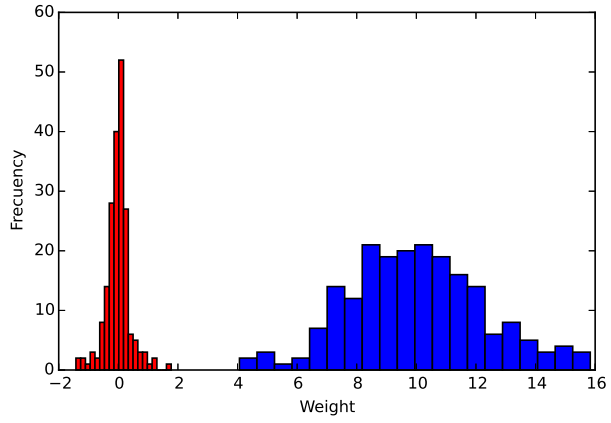

Fig. 1. Histograms of the distribution of the bootstrapped SVR weights for a relevant feature (blue) and an irrelevant feature (red). In this case 200 SVRs were trained.

IV. Solar radiance problem formulation

The problem we tackle in this paper, and the one we are using to evaluate the effectiveness of the ensemble selector algorithm, can be stated in the following way: Let $\mathcal{P}$ be a given point of the Earth's surface where the global solar radiation ($\mathcal{I}_t$) must be predicted ($\hat{\mathcal{I}}_t$), at a given time t. To do this, let us consider the output, $\mathcal{V}$, of a numerical meso-scale model $\mathcal{M}$, in a number M of nodes, consisting of the prediction at time t for N atmospheric variables, possibly at different heights, $\mathcal{V} = (\varphi_{11}, \ldots, \varphi_{1N}, \varphi_{21}, \ldots, \varphi_{2N}, \ldots, \varphi_{M1}, \ldots, \varphi_{MN})$, as shown in Figure 2.

This paper deals with the prediction of the global solar radiation registered in $\mathcal{P}$ at time t, using as predictive variables the set $\mathcal{V}$, or any subset of it. This type of problem is usually known in other works as a *statistical downscaling* for the solar radiation prediction of model $\mathcal{M}$ to point $\mathcal{P}$.

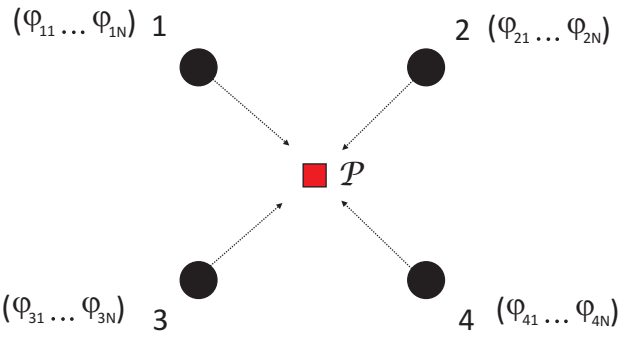

Fig. 2. Solar radiation prediction scheme used in this work for $M = 4$.

A. Model $\mathcal{M}$: the Weather Research and Forecasting model

In this case we use the well-known Weather Research and Forecasting (WRF) meso-scale model as $\mathcal{M}$ [5]. WRF is an extremely powerful meso-scale numerical weather prediction system designed for atmospheric research and also for operational forecasting needs. The WRF was developed in collaboration by the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA) of the USA. The WRF has been used in a wide range of meteorological [25] and renewable energy applications [26].

In our study, WRF model version 3.6 has been used. It has been run every 12 hours since it was started in 2011. Meteorological data are calculated over a window ranging in latitude from 34° 33' 43"N to 44° 28' 12"N , and in longitude from 4° 25' 12"W to 4° 23' 2"E. In this window, the grid has 99 elements from West to East, and 59 elements from North to South, roughly, each grid element covers 15×30 km^2. Atmospheric values are calculated, in the vertical dimension, at 37 levels above the ground, at ground level, and at four additional levels beneath the surface. The grid type is Arakawa, that is to say that data are calculated at the center of each element, with a 72 seconds time step.

WRF is initialized by data coming from NCEP FNL Operational Global Analysis and works in non-hydrostatic way. The short wave scheme used is that from MM5 shortwave (Dudhia), and the long wave model is the RRTM (Rapid Radiative Transfer Model). A radiation time step of 30 minutes was applied to each radiation domain. The land surface fluxes were obtained by Monin-Obukhov similarity theory, the surface physics was solved by the Unified Noah land surface model and the Planetary Boundary Level (PBL) by means of the Yonsei University (YSU) PBL scheme. The PBL was calculated at every basic time step, and five layers were considered in land surface model. Cumulus retrieval parameters was done by using the new Kain-Fritsch scheme, as in MM5 and Eta/NMM ensemble version, with a time step of 5 minutes.

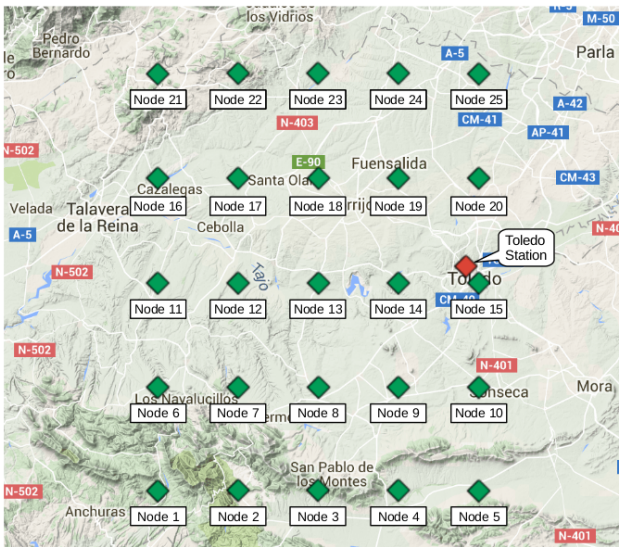

Fig. 3. Map showing the geographical locations of the 25 meso-scale nodes as well as the Toledo measurement station.

Finally, microphysics was carried out by the WSM 3-class scheme and the turbulent diffusion option was to select 2^{nd} order diffusion on model levels. This complements vertical diffusion done by the PBL scheme.

As objective variable data to train and test the algorithms, we consider the global solar radiation measured at Toledo's measuring station. Toledo station belongs to the measuring network of the Meteorological State Agency of Spain (AEMET), and it is located near the center of Spain (39° 53' 5"N, 4° 02' 43"W), at an altitude of 515 m. One complete year of hourly data (from May 1st, 2013 to April 30th, 2014) was used.

We use 25 nodes from the meso-scale model data as our problem input. Each node generates 423 variables including wind speed components and temperature values at different altitudes, upper atmosphere outgoing long wave radiation, cloud coverage per cell, etc. This results in a 25×423 sensor-variable matrix for each global solar radiation measurement. Figure 3 shows the geographical positions of the 25 meso-scale nodes and the Toledo measuring station.

A preliminary processing of the data reveals that many features present a standard deviation smaller than 0.01. We deemed these features irrelevant and subsequently eliminated them from the study. Overall, 301 node-variables survive this initial selection, resulting in a dimensionality reduction of around 30%. The resulting sensor-variable matrix has dimensions 25×301

V. Experiments and Results

A. Experimental setup

From the aforementioned model we obtain a data set consisting of 5840 samples with 7525 input features (resulting from the 25×301 sensor-variable matrix). For our experiments, the dataset is split into ten partitions that are used to calculate independent performance measurements. The measurements for the ten partitions are then averaged to obtain a single stable performance measurement for each of the methods under study.

To evaluate the BS-FS algorithm's performance we compare it to that of four standard feature selection algorithms as well as a baseline performance measure. The implemented methods are:

- A baseline method without feature selection to give us a reference performance value.
- A filter that ranks features in terms of their variance (variance filter). This is thus a non-supervised method. Features with higher variance are considered to be more relevant.
- A filter that ranks features in terms of their correlation with the target vector (correlation filter). Therefore, this is a supervised method. Features with higher correlation are considered to be more relevant.
- A Lasso regression model. The l_1 regularisation parameter was swept on a logarithmic scale between 10^{-5} & 10^{-2}.
- An RFE algorithm.
- The BS-FS algorithm with parameters $K = 100$ and $S = 0.5$ of the training data size.

All the methods described above are paired with an SVR trained with the resulting feature sets except for the Lasso method, since it is an embedded algorithm that trains its own regressor. This SVR uses a linear kernel and parameters (C = 1.0 and ϵ = 0.2, where C is the l_2 regularization coefficient and ϵ is the epsilon-tube coefficient). These parameters are fixed after analyzing their influence on performance and realizing that, for this database, there is a very flat region around these values where the SVR performs well. These same values are used inside the BS-FS and RFE algorithms for the internal SVR parameters.

B. Performance measurements

To evaluate the performance of these algorithms we use the R^2 metric, which is defined as:

$$R^2 = 1 - \frac{\sum_{l \in C_{tst}} (y_l - \hat{y}_l)^2}{\sum_{l \in C_{tst}} (y_l - \bar{y})^2}, \quad (3)$$

where y_l is the true target value for the l^{th} test sample, $\hat{y}_l$ is the predicted target value for the l^{th} test sample, $\bar{y}$ is the true mean of the test target values and C_{tst} is the set of test samples.

We then calculate the test curves for the different selection methods, which can be seen in Figure 4. These curves give us a good picture of how the features selected by the algorithms behave when they are fed to the final SVR.

Next, we validate the optimal number of features for the all the methods. This is done by using a 10-fold cross-validation strategy inside each of the ten data partitions. In the case of the BS-FS algorithm, It is important to note that the optimal number of features is associated with the optimal p-value for the hypothesis test. Table I shows the validation performances

TABLE I

FEATURE SELECTION PERFORMANCE ANALYSIS. ALL VALUES ARE OBTAINED BY CROSSVALIDATION OF THE WORKING POINT. ALL RESULTS ARE AVERAGED OVER THE 10 TEST PARTITIONS AND THEIR STANDARD DEVIATIONS ARE INCLUDED.

	Baseline	**Correlation**	**Variance**	**Lasso**	**RFE**	**BS-FS**
$\mathbf{R^2}$	0.922 ± 0.006	0.928 ± 0.004	0.923 ± 0.004	0.933 ± 0.004	0.926 ± 0.005	0.931 ± 0.004
Nº of features	All	1030 ± 603.41	270 ± 161.55	53 ± 2.25	1355 ± 140.46	450 ± 143.37

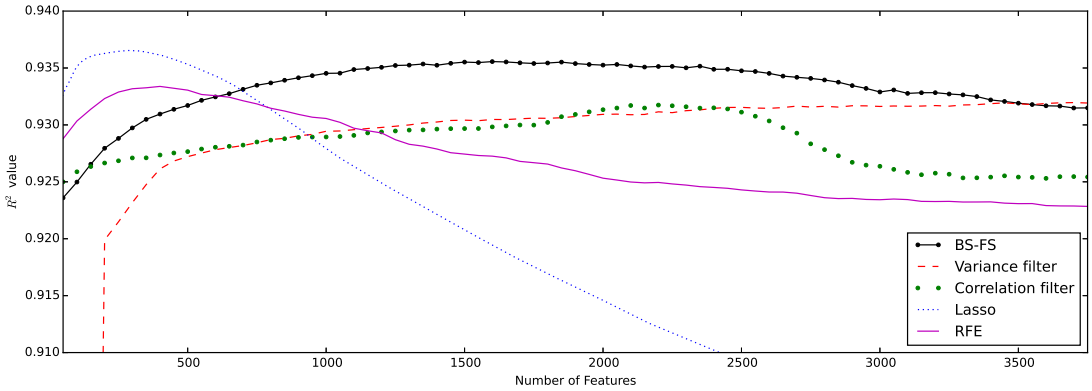

Fig. 4. Test R^2 evolution with respect to the number of selected features.

for all the methods in terms of the R^2 score and the number of selected features.

From these results we can see that feature selection provides an improvement not only in terms of complexity, since the number of features is considerably reduced in all cases, but also, albeit slightly, in terms of performance. In general, all the multivariant methods (BS-FS, Lasso and RFE) perform better than the univariant algorithms (correlation and variance filters).

Furthermore, from the test curves we can clearly see that both the RFE and Lasso present overfitting when using more than around 500 features. Moreover, their best performance is offered over a very narrow band of selected features. This leads to a negative impact in performance in the case of the RFE method, as can be seen in Table I. On the other hand, the BS-FS method presents a very flat maximum performance region that indicates its robustness against overfitting and its insensitiveness to fluctuations in the number of selected variables.

C. Interpretability analysis

We can also study the information that can be extracted from the selection provided by each of the methods. Figures 5, 6, 7, 8 & 9 represent the individual variable selection masks[3] averaged over the ten partitions arranged as a 25×301 node-variable matrix. This way, higher values (represented by hotter colors) indicate a higher consistency within a given method, whereas lower values (represented by cooler colors) show a lack of confidence in the variable's importance. Two bars, one below and another to the right of the matrix, indicate the consistency of global variables (understood as a variable considered over the 25 nodes) and nodes respectively. This consistency is calculated by averaging the number of individual variables, column-wise for global variables and row-wise for nodes, that present a consistency greater than 0.75. Again, higher values indicate greater consistencies and vice versa. This allows us to visualize the nodes and variables that each method considers important.

From the maps we can establish a method ranking in terms of how consistently selective and informative they are in both the node space and the variable space:

The worst performers in terms of interpretability are the RFE algorithm and the variance filter. Their selections are all over the map in both the sensor and variable spaces. Specifically, the variance filter is extremely noisy and inconsistent, giving us little to no information as to what variables or sensors may be the most helpful. The RFE fairs only slightly better, at least in the variable space, yet it still fails to provide useful information.

Next in line are the correlation filter and the Lasso. These methods seem to be able to be much more informative in the variable space, yet they present very noisy results in the node space. In the case of the correlation filter, very strong consistencies are found in a series of specific variable bands related to wind speed and temperature. The lasso tends to favor variables related to cloud coverage.

The BS-FS algorithm seems to be the most informative of all since it shows high consistencies in both the variable and

[3]The mask assigns a value of 1 if a variable is selected and of zero otherwise.

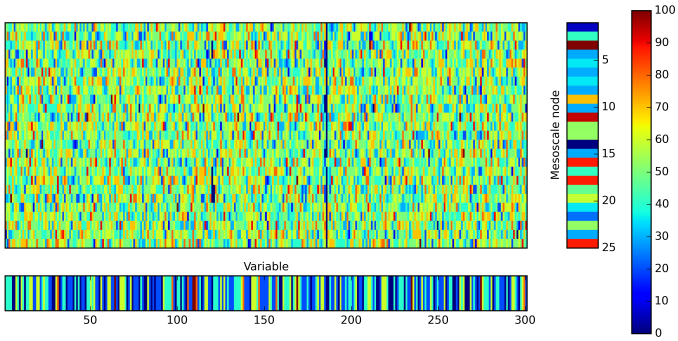

Fig. 5. Consistency map of the variables selected by the variance filter.

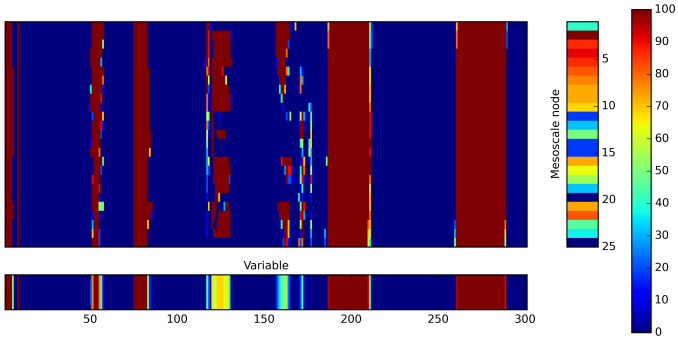

Fig. 6. Consistency map of the variables selected by the correlation filter.

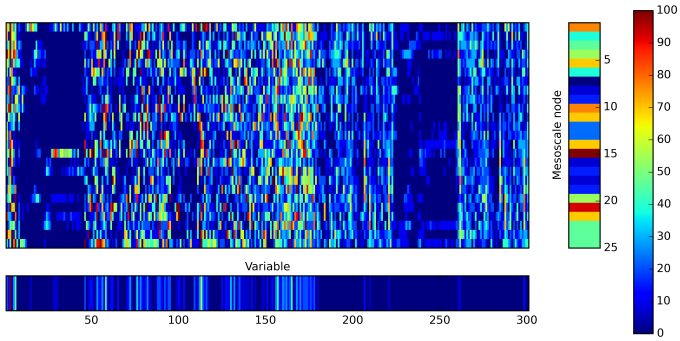

Fig. 7. Consistency map of the variables selected by the RFE algorithm.

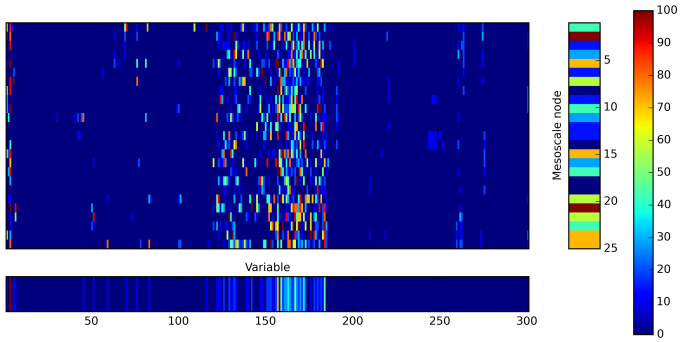

Fig. 8. Consistency map of the variables selected by the Lasso algorithm.

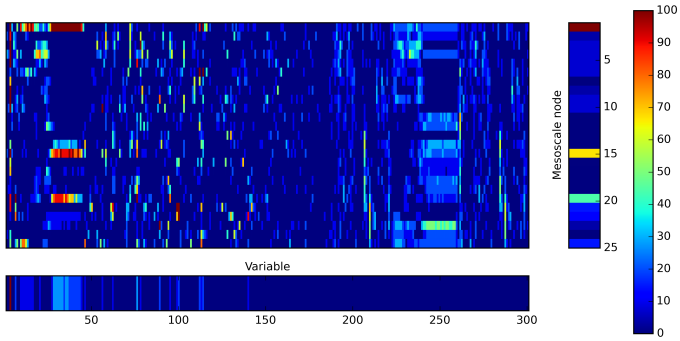

Fig. 9. Consistency map of the variables selected by the BS-FS algorithm.

node spaces. In the case of the global variables, this method places the most relevance on wind speed, temperature, atmospheric pressure and the downward short wave flux at ground surface. It also favors, albeit to a lesser degree, variables related to cloud coverage and rainfall. The most interesting information however is presented in the node space. It can be seen that three nodes are considered to be of particular importance: numbers 1, 15 & 25. The last two are interesting in the sense that they seem to confirm the soundness of the BS-FS's selection since they are after all the nodes that are closest to the Toledo measuring station (see figure 3). But Node 1 on the other hand is placed quite far from Toledo, and yet it is considered by the BS-FS algorithm to be the most important one. Interestingly, if we look at the geographical location of this node we can see that it lies over a tall hill that towers over the locations of all the other nodes. This could indicate a relationship between the quality of the information gathered by the meso-scale model and the altitude of the terrain that lies at the position of the nodes, particularly for the pressure and wind speed components.

VI. Conclusions

In this paper we present a novel feature selection method called Bootstrapped SVRs for Feature Selection (BS-FS) and we apply it to a solar radiation prediction problem. The BS-FS method aims at finding all the informative features by training a set of bootstrapped SVRs classifiers to analyze which classifier weights have been generated by a zero mean probability distribution and, therefore, are irrelevant for the regression task. We evaluate its performance with a Weather Research and Forecasting model consisting of 25 nodes, each of them containing a set of variables related to weather conditions, along one year (from May 1st, 2013 to April 30th, 2014) of hourly sampled data. Besides, its performance is compared with several well-known feature selection approaches: two fil-

tering schemes, the Recursive Feature Elimination method and the LASSO algorithm. Results corroborate the performance benefits of using feature selection in this learning task and point out the robustness of the proposed method, since it clearly alleviates the overfitting problems presented by most of the reference methods. Furthermore, when the maps of selected features are analyzed in detail, we realize that the proposed method is able to consistently locate global variables and nodes which are useful for solar radiation prediction in this particular case.

ACKNOWLEDGMENT

This work has been partially supported by the projects TIN2014-54583-C2-2-R and TEC2014-52289-R of the Spanish Ministerial Commission of Science and Technology (MICYT), and by Comunidad Autónoma de Madrid, under project PRICAM P2013ICE-2933.

REFERENCES

[1] T. Khatib, A. Mohamed, and K. Sopian, "A review of solar energy modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2864–2869, 2012.

[2] M. Bilgili and M. Ozoren, "Daily total global solar radiation modeling from several meteorological data," *Meteorology and Atmospheric Physics*, vol. 112, no. 3-4, pp. 125–138, 2011.

[3] S. Salcedo-Sanz, A. M Pérez-Bellido, E. Ortiz-García, A. Portilla-Figueras, and L. Prieto, "Hybridizing the fifth generation meso-scale model with artificial neural networks for short-term wind speed prediction," *Renewable Energy*, vol. 34, pp. 1451–1457, 2009.

[4] S. Salcedo-Sanz, A. M Pérez-Bellido, E. Ortiz-García, A. Portilla-Figueras, L. Prieto, and F. Correoso, "Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks," *Neurocomputing*, vol. 72, no. 4, pp. 1336–1341, 2009.

[5] W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, and J. Powers, "A description of the advanced research wrf version 2," tech. rep., National Center for Atmospheric Reserach, Mesoscale and Microscale Meteorology Division, 2005. Technical Note.

[6] M. Benghanem and A. Mellit, "Radial basis function network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at al-madinah, saudi arabia," *Energy*, vol. 35, no. 9, pp. 3751–3762, 2010.

[7] M. Behrang, E. Assareh, A. Ghanbarzadeh, and A. Noghrehabadi, "The potential of different artificial neural network (ann) techniques in daily global solar radiation modeling based on meteorological data," *Solar Energy*, vol. 84, no. 8, pp. 1468 – 1480, 2010.

[8] A. Linares-Rodriguez, J. Ruiz-Arias, D. Pozo-Vazquez, and J. Tovar Pescador, "An artificial neural network ensemble model for estimating global solar radiation from meteosat satellite images," *Energy*, vol. 61, pp. 636 – 645, 2013.

[9] C. Paoli, C. Voyant, M. Muselli, and M. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, no. 12, pp. 2146 – 2160, 2010.

[10] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 1–32, 1998.

[11] J. Chen, H. Liu, W. Wu, and D. Xie, "Estimation of monthly solar radiation from measured temperatures using support vector machines – a case study," *Renewable Energy*, vol. 36, no. 1, pp. 413 – 420, 2011.

[12] J. Chen, G. Li, B. Xiao, Z. Wen, M. Lv, C. Chen, Y. Jiang, X. Wang, , and S. Wu, "Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature," *Energy Conversion and Management*, vol. 89, pp. 318 – 329, 2015.

[13] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, and M. Sánchez-Girón, "Daily global solar radiation prediction based on a hybrid coral reefs optimization – extreme learning machine approach," *Solar Energy*, vol. 105, pp. 91 – 98, 2014.

[14] S. Shamshirband, K. Mohammadi, H. Chen, G. Samy, D. Petković, and C. Ma, "Daily global solar radiation prediction from air temperatures using kernel extreme learning machine: A case study for iran," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 134, pp. 109 – 117, 2015.

[15] L. Olatomiwa, S. Mekhilef, S. Shamshirband, and D. Petković, "Adaptive neuro-fuzzy approach for solar radiation prediction in nigeria," *Renewable and Sustainable Energy Reviews*, vol. 51, pp. 1784 – 1791, 2015.

[16] C. Voyant, M. Muselli, C. Paoli, and M. Nivet, "Hybrid methodology for hourly global radiation forecasting in mediterranean area," *Renewable Energy*, vol. 53, pp. 1 – 11, 2013.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song, "Dimensionality reduction via sparse support vector machines," *The Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[21] E. Parrado-Hernández, V. Gómez-Verdejo, M. Martínez-Ramón, J. Shawe-Taylor, P. Alonso, J. Pujol, J. M. Menchón, N. Cardoner, and C. Soriano-Mas, "Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction," *Medical image analysis*, vol. 18, no. 3, pp. 435–448, 2014.

[22] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[23] B. Efron, "Bootstrap methods: another look at the jackknife," *The annals of Statistics*, pp. 1–26, 1979.

[24] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[25] T. Giannaros, V. Kotroni, and K. Lagouvardos, "Predicting lightning activity in greece with the weather research and forecasting (wrf) model," *Atmospheric Research*, vol. 156, pp. 1 – 13, 2015.

[26] D. Carvalho, A. Rocha, M. Gómez-Gesteira, and C. Silva Santos, "Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the iberian peninsula," *Applied Energy*, vol. 135, pp. 234 – 246, 2014.