POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Multi-task Feature Selection for Drought Monitoring: a conditional mutual information approach

Master of Science in Computer Science and Engineering

Author: Lorenzo Nardi

Advisor: Prof. Marcello Restelli

Co-advisors: Prof. Andrea Castelletti, Prof. Matteo Giuliani, Dott. Alberto Maria Metelli, Dott. Paolo Bonetti

Academic year: 2021-2022

1. Introduction

The topic of climate change is particularly interesting nowadays, given that society is increasingly aware of the dramatic consequences to the ecosystem and human lives. Rising temperatures contribute to the increasing frequency of observing periods of drought, a phenomenon associated with the lack of water that impacts soil conditions for growing food, with strong economic and social consequences. Machine Learning is a branch of artificial intelligence that involves a series of tools and algorithms that are able to extract knowledge from past data, to gather insights on new samples. Over the years, various Machine Learning techniques have been employed to study and monitor prolonged drought periods over hydrological basins. FRIDA [1] is a data-driven approach to reconstructing a drought index from candidate drivers, which are selected through feature selection algorithms. In this thesis, we propose the extension of the FRIDA framework by employing an advanced feature selection technique based on the concept of Conditional Mutual Information and the Multi-Task Learning paradigm. In this way, we aim to generalize the application of a single model to evaluate drought conditions on a particularly extensive hydrological area, such as the Po Valley. The case study region is composed of several sub-basins, each having different geographical characteristics yet related by the presence of the Po River, which flows through them all. The research is conducted in two settings: Single-Task, where the chosen models are evaluated individually on hydrological sub-basins in the study area; Multi-Task, where a single model produces an output for each basin. The ultimate goal is to demonstrate that the Multi-Task model is able to generalize over the different areas under analysis, so that this approach can be reproduced over any geographic area.

2. Feature Selection

The growing popularity of the use of Machine Learning algorithms has led to the observation of increasingly large datasets, which constitute an obstacle to learning known as *curse of dimensionality*. Too many input variables and

few observations induce Machine Learning models to overfit, meaning that they are not able to achieve good performances on unseen data. Feature selection consists of techniques aimed at identifying the subset of non-redundant and relevant features, to learn a given task. Therefore, the objective is to reduce the dimension of the models inputs, which is helpful to ease the computational cost and improve the generalization capabilities. The features are discriminated using a choice criterion, that is applied to establish which of the available variables is relevant with respect to a target. In literature has been proposed broad categorization of available feature selection techniques. Typically, they are grouped in three categories [2]: *filters*, *wrappers*, *embedded methods*.

Wrapper methods evaluate the quality of the input variables based on the performance of a model. Therefore, every subset of features available is used to evaluate a model's predictions, and the process stops until some stopping criterion is met. Usually, this coincides when the highest prediction performance is obtained, or after it has been selected the desired number of variables. This feature selection approach is heavily dependent on the employed model class for evaluating the variables, and it has a high computational cost in the case of a large number of inputs.

Filter methods select features by ranking them according to a criterion. The ranking allows to identify the best inputs for a target and to define a threshold to select which one will be discarded. These techniques are the most efficient way to perform variable selection, as they do not involve the training of a model instance, and allow to obtain the best generalization capability because of exploiting significant relationships in the features. On the other hand, they usually get worse performance on specific models than the others because they are model agnostic.

Embedded methods are a combination of the previous approaches, that involve regularization techniques to speed up the learning phase of a model by filtering the features to be included in the evaluated subsets. Even though they are computationally faster than wrappers, the results of the variables selection still heavily depend on the model class used for evaluating the subsets.

In this thesis, a filter-type feature selection strategy is adopted since it allows to identify the subset of relevant and non-redundant variables with respect to the analyzed target, independently from the underlying model class. Specifically, the proposed approach relies on the concept of *Mutual Information* and *Conditional Mutual Information*, two measures from information theory that quantify the informativeness between a set of variables and a target, by capturing the nonlinear dependencies among them.

2.1. Mutual Information

Mutual Information (MI) measures how much information is possible to get about a random variable by observing another one.

The *entropy* H of a random variable X that has p as probability density function is:

$$\begin{aligned} H(x) =& \mathbb{E}_X[-\log(p(X))] \\ =& -\int p(x)\log p(x)dx, \end{aligned} \tag{1}$$

which describes the measure of uncertainty of the random variable X in terms of the probability of its occurrence. Introducing the *Kullback-Leibler divergence* (KL) of two distributions p and q as:

$$\begin{aligned} D_{KL}(p\|q) =& \mathbb{E}_{p(X)}\left[\frac{p(X)}{q(X)}\right] \\ =& \int p(x)\log\frac{p(x)}{q(x)}dx, \end{aligned} \tag{2}$$

then, we can define the Mutual Information between two random variables X and Y as it follows:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= D_{KL}(p(X,Y)||p(X)p(Y)) \\ &= \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy. \end{aligned} \tag{3}$$

The MI measures how much information the random variable X carries about Y, and vice versa. In feature selection terms, this allows to quantify the importance of a feature subset in relation to the target vector.

2.2. Conditional Mutual Information

Conditional Mutual Information (CMI) extends the MI by considering the occurrence of a third random variable Z. The CMI between X, Y giving the observation of Z is defined as:

$$I(X;Y|Z) = \mathbb{E}_Z[D_{KL}(p(X,Y|Z), p(X|Z)p(Y|Z))]$$

$$= \int p(z) \iint p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} dxdydz. \quad (4)$$

This measures how much information X carries about Y that is not already contained in Z. The novelty of the application of CMI in this thesis is to apply for the first time an algorithm of feature selection based on CMI [3] in the context of droughts monitoring.

2.3. Conditional Mutual Information based Feature Selection

The authors of the employed algorithm [3] provide that the theoretical learning errors are linked with conditional mutual information of the considered subset of features.
Specifically, feature selection is applied by leveraging greedy sequential search algorithms and imposing a stopping condition that guarantees that the ideal regression/classification error, obtained using the evaluated subset of features, stays below a user-defined threshold.

To select the best subset of features, we employ one of the two proposed methods in [3], thank to its faster computation time: the Forward selection algorithm. This algorithm adds to the result subset the feature that, at each step t, obtains the highest score in terms of CMI with respect to the target, conditioned on the features already included within the resulting set.
The proposed stopping conditions in [3] rely on a user-define parameter $\delta \geq 0$, which indicates the maximum error that a subset of features is allowed to introduce.

2.4. Estimation of Conditional Mutual Information

The concepts of MI and CMI can be computed exactly only on discrete variables. For this reason, when dealing with continuous values, there is a need for appropriate discretization techniques, which consequently do not allow an exact calculation of the mutual information value between two variables. In this thesis, we adopted the estimator proposed in [4], which allows to estimate mutual information between variables and targets of different nature (e.g., discrete target but continuous features) leveraging local properties identified through nearest neighbors. Indeed, we can define the CMI between X and Y, given Z, using the chain rule:

$$I(X;Y|Z) = I(X;Y,Z) - I(X;Z). \quad (5)$$

The estimator is then used to compute the individual terms that concur in the above equation.

3. Dataset

The proposed approach for drought monitoring is evaluated in two different settings:

1. a *regression* approach aimed at reconstructing the NDVI index through a combination of meteorological variables, which are easier to obtain than target satellite measurements, especially in the case the presence of clouds obscures the satellite images;
2. a multi-class *classification* problem, where the objective is to associate one sample of weekly climatological data to one among three identified labels describing a drought condition.

The target signal evaluated in this case study is the Normalized Difference Vegetation Index (NDVI), which can capture the health status of vegetation by assessing the amount of solar radiation absorbed by green leaves. NDVI data was extracted on the considered case study region, shown in Figure 1, hydrologically divided into ten main sub-basins.

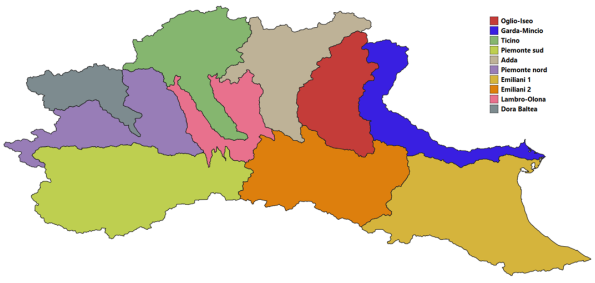

Figure 1: Study area considered for extraction of input and target variables.

The variables used as inputs for learning models are the following:

1. average temperature and precipitation in each of the sub-basins, since abrupt changes

of this phenomena can substantially damage crops;
2. average snow depth cover for every sub-basin (whenever applicable), that determines the availability of water supplies that accumulate during winter and get released in spring and summer;
3. average *lakes water levels*, because water shortage derived by droughts causes groundwater levels, streamflows and lake levels to reduce. Therefore, this acts as an indicator for prolonged water scarcity periods. In this case, we consider the major lakes that contribute to the water supply of areas of high interest, from both agricultural and tourism perspectives, which are: Como Lake, Iseo Lake, Maggiore Lake and Lugano Lake.

Since the extracted NDVI data are available as a weekly average, the features were also transformed to fit this scale. Then, increasing temperatures or reducing snow cover have a reasonably long-term impact on drought conditions. Thus, rather than exploring the link between weekly values of candidate features and the target, variables have been aggregated over $4, 8, 12, 16, 24$ weeks.

To make models learn the target peculiarity of a specific week in a given year, the feature selection process and then the experimental work is carried out by considering the NDVI anomaly as target. This measures how much the signal deviates from the average value it has in a given week, and it is helpful to remove the deterministic correlation of the target signals, due to the fact that NDVI has an increasing trend in summer, while a decreasing one during winter. The obtained NDVI anomalies show a positive linear trend with respect to temporal indicators. This is not desirable, since the employed methods assume a stationary target. For this reason, the anomaly signals are *detrended* by fitting a linear model on each and subtracting its values from the original series.
Also in the case of input data, we considered their anomalies, since it is reasonable to assume that anomalous behaviors of target signals should correspond to an anomaly in the observed features.

The dataset built in this way has an overall of 154 candidate features for 10 target variables. The considered time interval covers a period of almost 20 year of weekly NDVI anomalies values, that result in a dataset of dimension 1038×164. A quick inspection of the cross-correlation matrix of the inputs shows that metereological data over different regions are extremely correlated. This suggests that it is possible to aggregate variables across multiple areas. Dimensionality reduction has been performed by employing a *local* version of the PCA algorithm, keeping the different types of variables distinct and applying the component analysis on each of the defined aggregations (i.e., for every variable aggregated on $4, 8, 12, 16, 24$ weeks). Dimensionality reduction has been performed by employing a "local" version of the PCA algorithm, keeping the different types of variables distinct and applying the component analysis on each of the defined aggregations (i.e., for every variable aggregated on $4, 8, 12, 16, 24$ weeks). The produced datasets show way less cross-correlation among the inputs, even though we are sacrificing the geographical interpretability when looking at the selected features of each single basin. Nevertheless, the overall interpretability is not much affected, since we have preserved the groups of variables (i.e., snow, lakes, precipitations and temperature) as well as the temporal aggregations. Results in Figure 2 shows that in the Single-Task the first component of temperature and snow depth are selected as most informative variables for the majority of basins, while in the Multi-Task setting, a more varied selection of variables is performed.

4. Results

The experimental results have been obtained by feeding several learning models with the variables chosen by the Feature Selection based on CMI. The performances are evaluated in two different settings:

- *Single-Task*, where as the target area we considered the sub-basin named *Emiliani 2*, highlighted in orange Figure 1, because it appears to be the one containing the most significant percentage of cultivable areas for NDVI data extraction.
- *Multi-Task*, where the target is the vector of the NDVI anomalies across all the sub-

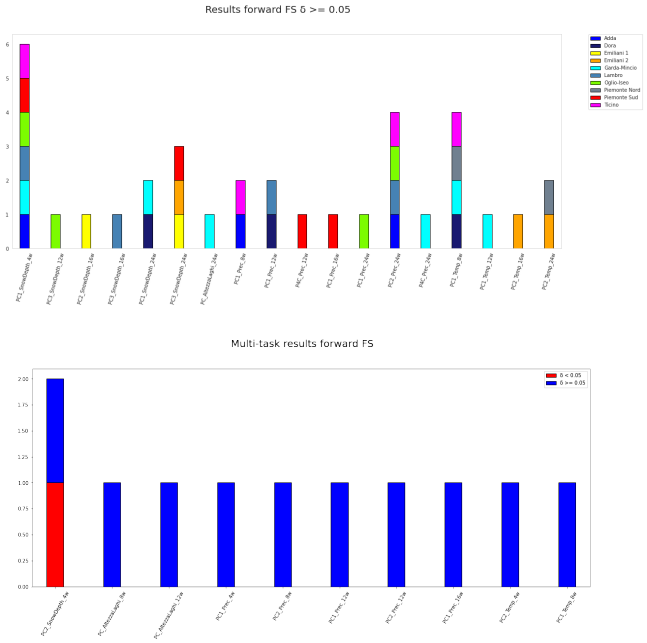

Figure 2: Forward selection results applied on the dataset resulting from PCA results.

basins in the study area.

To tackle the problem of noisy input measurements, the models behavior is investigated by considering past input values for up to 6 months. With this approach, we attempt to make the models find a relationship between inputs and the NDVI anomaly so that we are able to reproduce the drought index using only our extracted data. The dataset is divided into the following splits: 60% for training samples, and the remaining 40% equally divided between validation and test. Cross-validation for time-series is used to select the best model's hyperparameters, whenever applicable.

4.1. Single-Task Results

The first step in the Single-Task analysis is to verify if the NDVI anomaly signal is autoregressive, meaning that past values of the target are sufficient to predict its trend, and the selected features do not add any information to an autoregressive model (AR) prediction. This is achieved by assessing if the residuals of the AR predictions are white noise: in this case, the model cannot be further improved as all the pattern in the data have been already captured. The *Ljung-Box* test highlights that the AR of third order are indeed white noise, second phase in this setting focuses on obtaining an approximation of the NDVI anomaly signal using only the input variables chosen by the CMI feature selection algorithm. The models trained in this way could not fully capture the signal trend, except for the Recurrent Network. In fact, the model output demonstrated to be able to reconstruct the target sign in most cases, an intuition that led us to consider a classification problem. We identified three conditions in which the anomaly of the NDVI can be found, which describe as many drought conditions:

- *Normal* indicates that the anomaly value is average; therefore, no drought conditions are detected.
- *Good* is assigned to anomaly values above average, which indicates and is an indicator of healthy vegetation.
- *Bad* identifies the situation in which anomaly values are below the average for a given week, meaning that there is a water shortage and, consequently, a drought condition.

The resulting class labels are balanced by construction in the training set, so that it is not necessary to employ imbalanced classification approaches. In this setting, the models are able to perform better in recognizing a drought condition, which is encoded into one of the labels identified from the target signal. The analysis shows that complex neural network models obtain discrete performances even if they suffer from few available data and noisy input values. To mitigate the first problem, the models are again given in input the historical data of the selected features. The results in Figure 3 show that the Feed-Forward network performs well in recognizing the target signal good conditions. As an additional indicator of robustness, the predictions are kept stable for consecutive weeks even though the model has no indicator of time.

4.2. Multi-Task Results

Starting from the good results of the Single-Task classification, the analysis in the Multi-Task setting is performed by training the models using the historical data of the selected features. Although the number of selected variables is higher than in the Single-Task setting, only the first 5 variables are kept, whose CMI scores are larger by at least an order of magnitude than the re-

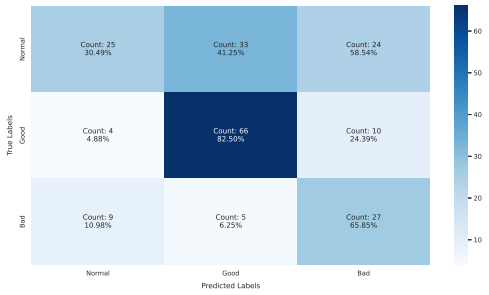

Figure 3: Confusion matrix of the Single-Task Feed-Forward Network.

maining. In this way, it is possible to consider the series of past input values again, without nullifying the dimensionality reduction obtained by the PCA algorithm.

Comparing the accuracy of predictions over the same area used in the Single-Task case, a deterioration in performance is observed. However, the models in this setting achieve discrete performance also on other areas, allowing to have satisfactory results training one single neural network for all the ten basins.

Two different approaches [5] were employed to enable knowledge sharing between tasks: the first one achieves this by sharing a part of the model architecture; the second one assigns shared layers to tasks grouped in clusters, based on their MI scores. All the experimental tests are carried out by using only Neural Networks models, since their architecture is more adaptable to enable the learning of a common representation for the input features. Evaluating the performance obtained in the two different cases of Multi-Task learning approaches, the model that seems most promising is the Feed-Forward Neural Network having a single hidden layer shared by multiple targets, whose performance on the sub-basins is shown in Figure 4. Unlike the Single-Task Feed-Forward model, which keeps the predicted labels stable for consecutive weeks, in this case the model seems to change more frequently the predicted label for a sample. Nevertheless, the task clustering did not benefit when applied to this type of network. Indeed, this approach does not consider the likely negative cross-correlations between tasks belonging to different clusters, which may result in updating the model parameters differently. In contrast, in the case of the Recurrent Neural Net-

Figure 4: Confusion matrices of the Multi-Task Feed-Forward Network

work, grouping targets result in a boost in predictive performance across the considered areas. Despite the deterioration of predictions in the transition to the Multi-Task case, this analysis verified that this approach still yields decent results in the presence of noisy input measurements and few data. In addition, this Multi-Task framework allows to analyze multiple areas in parallel to assess their drought conditions, with the only need to train a single neural network.

5. Conclusions

The goal of this thesis is to enhance Drought Monitoring employing an advanced feature selection technique, based on the concept of Conditional Mutual Information, and Multi-Task models for capturing common drought drivers across different geographical areas. The experimental analysis is carried out by using meteorological data on ten sub-basins identified in the study area, the Po Valley. The dimensionality of the dataset is then reduced with a local version of the PCA algorithm, so to keep the nature and the temporal aggregation of the extracted variables, that are originally highly correlated. The models performances are evaluated in two settings: the classical Single-Task approach, and the Multi-Task paradigm. Ini-

tially, we discovered that the NDVI anomaly signal is autoregressive, then we focused on trying to obtain a good approximation of the target using only the chosen features. However, the reconstructed signal appeared lagged with respect to the original data and not enough accurate, most likely due to the noisy values the model received in input. Nevertheless, the Recurrent Neural Network prediction trend has shown that the model has learned to recognize the sign of the target signal, which prompted us to try a classification approach. The objective is to assign each sample one among three labels, representing three different conditions of the signal, which we have interpreted with as many drought conditions. This setting led to positive results, since the chosen models were able to extrapolate information from the input data, in a context with limited number of samples and noisy data obtaining good performances on one evaluated area. Although the results in the Multi-Task case seem to be worse when comparing the prediction accuracy on the same area used in the Single-Task analysis, the overall performance shows that the models were able to generalize even with few data. This suggests that this setting may be particularly interesting for predicting future drought conditions on multiple geographically distinct areas in parallel.

References

[1] Marta Zaniolo, Matteo Giuliani, Andrea Francesco Castelletti, and Manuel Pulido-Velazquez. Automatic design of basin-specific drought indexes for highly regulated water systems. *Hydrology and Earth System Sciences*, 22(4):2409–2424, 2018.

[2] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[3] Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni, and Marcello Restelli. Feature selection via mutual information: New theoretical insights. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019.

[4] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30, 2017.

[5] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.