ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Spatio-temporal feature selection for black-box weather forecasting

Zahra Karevan and Johan A. K. Suykens

KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Abstract. In this paper, a data-driven modeling technique is proposed for temperature forecasting. Due to the high dimensionality, LASSO is used as feature selection approach. Considering spatio-temporal structure of the weather dataset, first LASSO is applied in a spatial and temporal scenario, independently. Next, a feature is included in the model if it is selected by both. Finally, Least Squares Support Vector Machines (LS-SVM) regression is used to learn the model. The experimental results show that spatio-temporal LASSO improves the performance and is competitive with the state-of-the-art methods. As a case study, the prediction of the temperature in Brussels is considered.

1 Introduction

Reliable weather forecasting is one of the challenges in climate informatics. It includes accurate predictions for many variables like temperature, wind speed, humidity, and precipitation. State-of-the-art methods use Numerical Weather Prediction which is a computationally intense method [1]. Recently, data-driven models have been used in weather and climate science to have insight into the embedded knowledge. There are different types of data-driven methods that have been used for weather forecasting both in linear and nonlinear frameworks. Two of the most popular methods are Artificial Neural Networks (ANN) and Least Squares Support Vector Machines (LS-SVM). In [2], the authors claim that LS-SVM generally outperforms artificial neural networks. Moreover, the effectiveness of LS-SVM for temperature prediction is demonstrated in our previous works [3, 4].

Weather forecasting can be considered as a time-series problem. Therefore, in order to have a reliable prediction for one specific day, not only the variables of the day before, but also some previous days are included in the prediction model [4]. Having several weather variables available for some locations and for several days leads to a large feature vector size and hence feature selection becomes of great interest to decrease the complexity of the model. In our previous work [4], a combination of k-Nearest Neighbor and Elastic net is used to reduce the number of features.

Furthermore, historical weather data can be considered as spatio-temporal data since they involve both place and time in the records [5]. In [6], Spatio-temporal Relational Random Forest is proposed to predict severe weather phenomena such as tornado and drought. The authors in [7] present a weather forecasting model based on exploring the joint influence of weather elements with considering spatio-temporal structure of the data.

In this paper, a feature selection method is proposed by taking spatio-temporal properties of the dataset into account. The main method for feature selection is LASSO [8] which is a well-known approach and is a penalized least squares method imposing an L_1-penalty on the regression coefficients. It is applied on the spatial and temporal parts of the dataset independently and the relevant features are selected based on the selected variables shared in both models. Finally, in order to model the data, Least Squares Support Vector Machines (LS-SVM) [9], which use a set of linear equations to solve the optimization problem, is used.

2 Background

In this section, LASSO and LS-SVM are explained. The former is used to to decrease the number of the features in the model and the latter is used as regression method.

2.1 LASSO

Due to the high dimensionality of the dataset, the feature selection is an important step in obtaining the relevant features. In this study, LASSO [8] is used as a feature selection approach. Let x be the feature vector and $x_{(i)}$ be the ith feature. Consider the following linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{(1)} + \ldots + \hat{\beta}_d x_{(d)} \, . \tag{1}$$

Several methods have been proposed for model fitting which lead to an estimation of β values. One of the popular ones is LASSO which is a regularization method that penalizes least squares imposing an L_1-penalty on the regression coefficients. Assume that there is a dataset with N observations and d variables. Let $y = [y_1, y_2, \ldots, y_N]^T$ and $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d \times N}$ where x_j is a vector including d features and y_j is the response value at observation j. LASSO solves

$$\hat{\beta} = \arg\min_{\beta} ||y - X^T\beta||^2 + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{2}$$

Because of the L_1-penalization, many of the coefficients shrink to zero and as a result a sparse model is produced. Note that in (2) λ is a positive tuning parameter.

2.2 Least Squares Support Vector Machines

Least Squares Support Vector Machines (LS-SVMs) [9], are used as a regression method and results in solving a set of linear equations. Let $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ and $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^h$ where $\varphi(\cdot)$ is a mapping function to a high or infinite dimensional space (feature map). The model in primal space is formulated as:

$$\hat{y} = w^T \varphi(x) + b \tag{3}$$

where $b \in \mathbb{R}$ and the dimension of w depends on the feature map and is equal to h. Let $\{x_j, y_j\}_{j=1}^{N}$ be the given training set, γ be the positive regularization

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

parameter and e_j be the error between the actual and predicted output for sample j equal to $e_j = y_j - \hat{y}_j$. The primal problem is given by [9]

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{j=1}^{N} e_j^2, \text{ subject to } y_j = w^T \varphi(x_j) + b + e_j, j = 1, ..., N. \tag{4}$$

It is obvious that if w is infinite dimensional, this optimization can not be solved in primal space; thus, the problem is solved in dual space. Assuming $\alpha_j \in \mathbb{R}$ are the Lagrange multipliers, from the Lagrangian $\mathcal{L}(w, b, e; \alpha) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{j=1}^{N} e_j^2 - \sum_{j=1}^{N} \alpha_j (w^T \varphi(x_j) + b + e_j - y_j)$, the LS-SVM model as a function estimator is obtained as follows

$$\hat{y} = \sum_{j=1}^{N} \alpha_j K(x, x_j) + b \tag{5}$$

where $K(x_j, x_l) = \varphi(x_j)^T \varphi(x_l)$ for $j, l = 1, 2, \ldots, N$ based on Mercer's theorem [10]. In this paper, the Radial Basis Function (RBF), in which $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2 / \sigma^2)$, is used as a kernel function. In this case, the regularization parameter γ and the kernel parameter σ are tuning parameters.

3 Spatio-Temporal LASSO

In data-driven weather prediction the features are historical weather variables. Looking into weather forecasting as time-series problem, the weather can be forecasted by a Nonlinear AutoRegressive eXogenous (NARX) model taking into account some historical data of different weather stations. It is obvious that each variable is recorded in the specific time and place and as a result weather data sets have spatio-temporal structure. In addition, having several weather elements for each day in a NARX model causes a high dimensional feature vector and feature selection is needed to find the relevant features. In our previous work [4], LASSO and Elastic net were used to reduce the number of features. In this paper, LASSO is utilized as feature selection method, while the spatio-temporal structure of data is taken into consideration.

Let $y(t)$ be the target value at time t and $X_q(t) \in \mathbb{R}^{d \times N}$ is a vector including all of the features at time t for the qth city. Assume $X_{spatial}(q) = [X_q(t-1); X_q(t-2); \ldots; X_q(t-lag)] \in \mathbb{R}^{d' \times N}$ and $X_{temporal}(p) = [X_1(t-p); X_2(t-p); \ldots; X_Q(t-p)] \in \mathbb{R}^{d'' \times N}$ where $q \in \{1, 2, \ldots, Q\}$ and $p \in \{1, ..., Lag\}$ where $d' = d \times Lag$ and $d'' = d \times Q$ and d is the number of measured weather elements in one day. Note that Q is the total number of cities and Lag is the number of previous days in the historical data used in the forecasting task.

In other words, each observation in $X_{spatial}(q)$ includes the historical weather elements for 1 to p previous days of the city q and $X_{temporal}(p)$ includes the historical weather elements for the p-th previous day for all Q cities. Consequently, for each city, there is a $X_{spatial}(q)$ and for each previous days there is a $X_{temporal}(p)$. Therefore, it can be concluded that historical weather elements are included in both $X_{spatial}$ and $X_{temporal}$. For example, the real measurements of 7 days ago in city 1 are present both in $X_{spatial}(1)$ and $X_{temporal}(7)$.

After creating $X_{spatial}$ and $X_{temporal}$ for different cities and Lags, LASSO is used to reduce the number of features both in the spatial and temporal direction, independently. In (6) and (7) and (6), the optimization problem in the spatial and temporal scenarios are shown respectively:

$$\hat{\beta}_{(q)} = \arg\min_{\beta_{(q)}} ||y - X_{spatial}(q)^T \beta_{(q)}||^2 + \lambda \sum_{j=1}^{d'} |\beta_{(q)j}|, q \in \{1, 2, ..., Q\}, \quad (6)$$

$$\hat{\beta}_{(p)} = \arg\min_{\beta_{(p)}} ||y - X_{temporal}(p)^T \beta_{(p)}||^2 + \lambda \sum_{j=1}^{d''} |\beta_{(p)j}|, p \in \{1, 2, ..., Lag\}. \quad (7)$$

It is obvious that the total number of linear models created by LASSO is $Lag+Q$ and $\hat{\beta}_{(q)} \in \mathbb{R}^{d'}$ and $\hat{\beta}_{(p)} \in \mathbb{R}^{d''}$.

For feature selection using LASSO, the variables with non-zero coefficients are considered to be relevant and are selected. In the proposed method, a feature is selected only if it is considered relevant both in the spatial and temporal parts. Thus, the particular feature in p days ago in city q is a relevant one if both coefficients in $\hat{\beta}_{(p)}$ and $\hat{\beta}_{(q)}$ are non-zero.

After finding important features, the feature vector includes all of the selected variables for Q cities and in Lag previous days. Afterwards, a LS-SVM model in trained as function estimator and used for forecasting.

4 Experiments

In this study, data are collected from the weather underground website which is one of the popular ones in weather forecasting. The data include real measurements for weather elements such as minimum and maximum temperature, dew point, precipitation, humidity and wind speed from the beginning of 2007 until mid 2014 and for 10 cities including Brussels, Liege, Antwerp, Amsterdam, Eindhoven, Dortmund, London, Frankfurt, Groningen and Dublin.

As in our previous work[4], in order to evaluate the performance of the proposed method, the experiments are conducted on two different test sets: one from mid-November 2013 until mid-December 2013 (testset1) and the other one from mid-April 2014 to mid-May 2014 (testset2). The prediction is done on daily basis and for each test set, the training data includes daily weather variables of all of the 10 cities from the beginning of 2007 until the previous day of the test set. Thus, the total number of samples in training set is about 2500. Also, the number of measured weather elements for each day is equal to 18.

To have a good generalization performance, all of the parameters are tuned using 10-fold crossvalidation. "tunelssvm" function in the LS-SVMlab1.8 is used for tuning γ and σ, the "lasso" function of MATLAB is used for tuning λ. Also, considering the problem as time-series one, the Lag variable is tuned by grid search in the range of 7 to 16. The performance is evaluated based on Mean Absolute Error (MAE) of the predictions.

In Figure 1 the performance of weather underground predictions for the minimum and maximum temperature in Brussels for both test sets together, is compared with LS-SVM without feature selection and the cases that LASSO and spatio-temporal LASSO are applied. As it is shown, the performance of the proposed method for the minimum and maximum temperature prediction mostly outperforms the cases where there is no feature selection or LASSO is applied while ignoring the spatio-temporal structure of data.

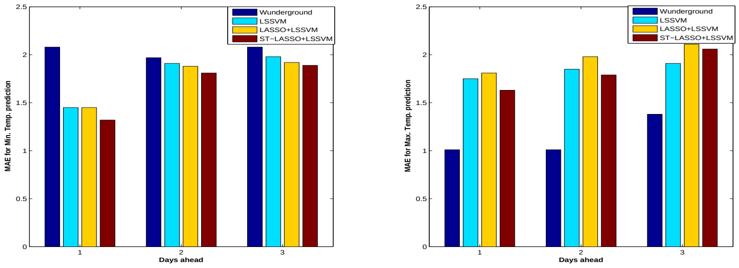

Fig. 1: MAE of the predictions in weather underground, LS-SVM, LASSO+LS-SVM and ST-LASSO+LS-SVM

Data set	Days ahead	Temp.	WU	LS-SVM	LASSO + LS-SVM	STLASSO + LS-SVM
Testset1	1	Min	1.57	1.49±0.005	1.51±0.001	**1.33**±0.009
		Max	**0.96**	1.23±0.003	1.39±0.005	1.16±0.005
	2	Min	**1.57**	1.61±0.001	1.87±0.02	1.75±0.005
		Max	**1.15**	1.39±0.005	1.49±0.01	1.37±0.01
	3	Min	1.76	**1.66**±0.005	1.91±0.05	1.82±0.02
		Max	**1.26**	1.50±0.001	1.79±0.01	1.81±0.02
Testset2	1	Min	2.59	1.38±0.005	1.46±0.005	**1.33**±0.01
		Max	**1.07**	2.21±0.004	2.21±0.003	**2.11**±0.005
	2	Min	2.37	2.01±0.004	1.98±0.02	**1.85**±0.01
		Max	**0.88**	2.25±0.005	2.31±0.05	2.25±0.01
	3	Min	2.40	2.02±0.002	2.09±0.03	**2.01**±0.02
		Max	**1.51**	2.40±0.005	2.51±0.05	2.44±0.05

Table 1: MAE and its variance of the predictions in weather underground (WU), LS-SVM, LASSO+LS-SVM and ST-LASSO+LS-SVM in testset1 (Nov/Dec) and testset2 (Apr/May).

The average MAE of the methods on each testset for 5 iteration can be found in Table 1. It can be observed that for the minimum temperature, the data-driven approaches mostly outperform weather underground company. Among the data-driven methods, the proposed method mostly has a better performance. This means that with the help of the spatio-temporal structure, better features are selected. In the experiments, it was observed that while preserving the sparsity, the average number of features selected by the proposed method is larger than the number of features selected by LASSO.

5 Conclusion

In this paper, a spatio-temporal LASSO feature selection for weather prediction was proposed. It mostly outperforms LASSO by incorporating the spatio-temporal structure of the data. The performance is analyzed by minimum and maximum temperature forecasting in Brussels in two different time periods for 1 to 3 days ahead.

Acknowledgment

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Research Council KUL: CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grant iMinds Medical Information Technologies SBO 2015 IWT: POM II SBO 100031 Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017).

References

[1] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[2] A. Mellit, A. Massi Pavan, and M. Benghanem. Least squares support vector machine for short-term prediction of meteorological time series. *Theoretical and applied climatology*, 111(1-2):297–307, 2013.

[3] Marco Signoretto, Emanuele Frandi, Zahra Karevan, and Johan A. K. Suykens. High level high performance computing for multitask learning of time-varying models. *IEEE Symposium on Computational Intelligence in Big Data*, 2014.

[4] Zahra Karevan, Siamak Mehrkanoon, and Johan A. K. Suykens. Black-box modeling for temperature prediction in weather forecasting. In *International Joint Conference on Neural Networks*, 2015.

[5] Noel Cressie and Christopher K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.

[6] A. McGovern, T. Supinie, D. J. Gagne II, N. Troutman, M. Collier, R. A. Brown, J. Basara, and J. K. Williams. Understanding severe weather processes through spatiotemporal relational random forests. In *NASA conference on intelligent data understanding*, 2010.

[7] Aditya Grover, Ashish Kapoor, and Eric Horvitz. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–386. ACM, 2015.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[9] Johan A. K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least squares support vector machines*. World Scientific, 2002.

[10] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, pages 415–446, 1909.