Black-box modeling for temperature prediction in weather forecasting

Zahra Karevan
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email:zahra.karevan@esat.kuleuven.be

Siamak Mehrkanoon
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email:siamak.mehrkanoon@esat.kuleuven.be

Johan A.K. Suykens
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email:johan.suykens@esat.kuleuven.be

***Abstract*—Accurate weather forecasting is one of most challenging tasks that deals with a large amount of observations and features. In this paper, a black-box modeling technique is proposed for temperature forecasting. Due to the high dimensionality of data, feature selection is done in two steps with k-Nearest Neighbors and Elastic net. Next, Least Squares Support Vector Machine regression is applied to generate the forecasting model. In the experimental results, the influence of each part of this procedure on the performance is investigated and compared with "Weather underground" results. For the case study, the prediction of the temperature in Brussels is considered. It is shown that black-box modeling has a good and competitive accuracy with current state-of-the-art methods for temperature prediction.**

I. Introduction

Accurate weather forecasting is one of the complicated and challenging tasks for climate researchers. It includes predicting many variables like temperature, wind speed, humidity, and precipitation. However, there are many parameters that make ideal weather forecasting barely possible, e.g. environmental issues like topography and surrounding structures of a particular place and chaotic characteristics of the atmosphere, or some other factors such as the shortage of knowledge about the atmospheric process or the influence of human on climate changes.

Recently, data-driven models have been used in weather and climate science and aim toward having insight into imbedded knowledge. Because of the insufficient reliable real measurements for weather elements, previous works looked into the weather forecasting by considering only local data, e.g. in [1], [2] and [3] or use globally simulated datasets, e.g. in [4]. Researchers use different types of data-driven methods to predict for different weather elements such as precipitation, speed of the wind etc. [5]–[8]. There have been some methods proposed for temperature prediction: in [3] an ensemble of connectionist learning methods was proposed to predict temperature and wind speed. The authors claimed that this method outperforms the multilayer perceptron, RBF and Elman recurrent neural network. Prediction for hourly temperature using neural networks was investigated in [2]. According to the authors, the model only uses the temperature feature of a particular time in the previous day. In [9], a neural network method was proposed for temperature prediction at 11 locations; the model uses a lattice of surrounding locations as input. In [10], as a preprocessing step, feature selection for the multilayer perceptron was used and afterwards a hybrid neural network using multilayer perceptron networks and self-organizing maps was proposed. In [4] sparse group LASSO was used for prediction. In [11], LS-SVM was used for the minimum and maximum temperature prediction in 350 stations and in [12], the authors claim that LS-SVM generally shows better performance rather than artificial neural networks.

One of the aims of this paper is to use real measurement values of weather elements instead of simulated data sets. The problem however is that it is difficult to capture reliable features in many weather stations at every time spot. Hence, there are often missing values in many features of some weather stations. There are different proposed methods for dealing with this issue, e.g. in [11], features with lots of missing values are eliminated from the data of each station. Moreover, it is seen that for some of the days, all the weather features are missing. Therefore, the days in the data set are not sequential. To overcome this problem, in [1] linear interpolation was used to estimate missing values in features. As a result, in many of the mentioned papers, only few variables for the weather stations had real values and were used as the input of the model.

In this paper, a consistent feature set for all of the stations and for all days in the particular time period, is taken into consideration. The real measured values of features are gathered from the weather underground website[1]. In addition, more weather variables like direction of the wind, snow, rain, fog, minimum, mean and maximum values for temperature, wind speed, humidity and sea level pressure, are incorporated for weather forecasting. In [3], it is stated that the temperature shows a high correlation with other weather elements and depends on many factors and even a small error in them can decrease the prediction accuracy significantly. Besides, it can be measured with higher degree of accuracy than other variables. Therefore, the main focus of this study is on predicting the minimum and maximum temperature.

In this study, the data are collected from 70 stations on the globe, mostly located in North America, East Asia and Europe. The diversity in the location of the stations makes it possible to compare the performance of considering data globally or locally. In favor of localizing, k-NN (k-Nearest Neighbor) is used assuming all of the data for each city as a sample. k-NN finds the most similar cities to the target one in terms of similarity in the weather elements over time and regardless of their geographical locations.

[1] www.weatherunderground.com

In order to have a more accurate model for weather forecasting, it is considered as a time-series problem. Therefore, the weather elements in one particular day are not only dependent on the day before, but also some previous days are taken into account for prediction. Having the features of many stations for several days results into a large feature vector and there is an absolute need for feature selection. In this paper, Elastic net, which is a combination of L_1-norm and L_2-norm, is used as the feature selection method. Note that if the part which is related to the L_2-norm is ignored, Elastic net represents LASSO [13] which is popular for making the model sparse. On the other hand, if the L_1-norm is disregarded, it corresponds to ridge regression [14]. In this study, Least Squares Support Vector Machines (LS-SVM) [15], which is a powerful machine learning method, are used for modeling. In comparison with SVM, it uses a set of linear equations, instead of convex quadratic programming, to solve the optimization problem.

This paper is organized in three sections: first the main components of the proposed method are described. Then, in the second section, these elements are attached together and the proposed method is explained and finally experimental results compared with one of the high quality forecasting companies (weather underground) predictions.

II. USED METHODS

In this section, k-Nearest Neighbor, Elastic net and LS-SVM are reviewed. The first two are used to produce a sparse model and consequently decrease the complexity in the proposed method. The latter is used as regression method. The choice of these methods is because of their success in a wide range of problems.

A. k-Nearest Neighbors

Given a set of observations in d-dimensional space, finding nearest neighbors is discovering the closest samples to a particular sample based on a similarity criterion. Several methods have been proposed to find the similarity between a pair of samples. In this paper, the evaluation of similarity between two observations is calculated based on the RBF kernel [16] as follows:

$$K(x_i, x_j) = \exp(-||x_i - x_j||_2^2/\sigma^2) \tag{1}$$

where σ is the kernel parameter and $x_i, x_j \in \mathbb{R}^d$. There are many techniques for tuning the kernel parameter σ; many of them are computationally expensive [17], [18]. Here the Silverman's Rule of Thumb [19] is used to select the kernel bandwidth. Assuming $\sigma \in \mathbb{R}^+$ and $\hat{\sigma}$ be the mean standard deviation for all features, then, using this rule, an estimation of σ value can be achieved as follows:

$$\sigma = \hat{\sigma} N^{(-1/(d+4))} \tag{2}$$

where N and d are the number of samples and features respectively. Then, the kernel matrix which includes similarity values for each pair of cities, can easily be constructed by

$$\Omega = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_N, x_1) \\ \vdots & \ddots & \vdots \\ K(x_1, x_N) & \cdots & K(x_N, x_N) \end{pmatrix}. \tag{3}$$

Afterwards, the k most similar samples to x_i are obtained by finding the k largest values in the ith column in the kernel matrix and choosing the corresponding samples.

B. Elastic net

Due to the high dimensionality of the dataset, the feature selection is an important step in obtaining the relevant features. Here, Elastic net [20] and LASSO [13] are used to reduce the number of features. Let x be the feature vector and $x_{(i)}$ be the ith feature. Consider the following linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{(1)} + \ldots + \hat{\beta}_d x_{(d)}\,. \tag{4}$$

Several methods have been proposed for model fitting which lead to an estimation of β values. LASSO is a well-known feature selection approach which is a penalized least squares method imposing an L_1-penalty on the regression coefficients. Beside continuous shrinkage, LASSO produces a sparse model and is being used as feature selection method. However, it has its own limitations. For example, it is unable to show the grouping information. If there are some features which are highly correlated, LASSO tends to choose only one of them, no matter which one. In addition, if the number of features is larger than the number of samples, LASSO cannot select more features than the number of observations. In order to avoid these limitations, Elastic net is employed. Elastic net is another optimization method for model fitting which benefits from the LASSO advantages and also has the ability to reveal the grouping information.

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_i and y_i are a vector including d features and the response value at observation i respectively. Elastic net solves

$$\hat{\beta} = \arg\min_{\beta} J(\beta, \lambda_1, \lambda_2), \tag{5}$$

where

$$J(\beta, \lambda_1, \lambda_2) = ||y - X^T\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||^2,$$

$$||\beta||^2 = \textstyle\sum_{j=1}^{p} \beta_j^2, \tag{6}$$

$$||\beta||_1 = \textstyle\sum_{j=1}^{p} |\beta_j|.$$

Note that λ_1 and λ_2 are penalty parameters. Assuming $\nu = \lambda_2/(\lambda_2 + \lambda_1)$ then the Elastic net minimization is in an equivalent form of

$$\hat{\beta} = \arg\min_{\beta} ||y - X^T\beta||^2, \tag{7}$$

$$\text{subject to } (1-\nu)||\beta||_1 + \nu||\beta||^2 \leq \eta;\ \textit{for some } \eta.$$

The term $(1-\nu)||\beta||_1 + \nu||\beta||^2$ is called Elastic net penalty which is a convex combination of L_1-norm and L_2-norm. For the value of $\nu = 1$, the optimization formula becomes ridge regression [14], while for $\nu = 0$, it represents LASSO. In this paper, it is assumed that $\nu \in [0, 1)$.

Experiments on real world datasets show that in addition to sparse representation advantages, Elastic net usually outperforms LASSO in terms of accuracy if the number of features is much larger than the number of samples.

C. Least Squares Support Vector Machines

In this paper, Least Squares Support vector Machines (LS-SVMs), proposed in [15] [21], are used as machine learning method and results in solving a set of linear equations. Let $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ and $\varphi : \mathbb{R}^d \to \mathbb{R}^h$ where $\varphi(\cdot)$ is a mapping function to a high or infinite dimensional space (feature map). The model in primal space is formulated as:

$$y(x) = w^T \varphi(x) + b \tag{8}$$

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 training set, γ be regularization parameter and e_j be the error between the actual and predicted output for sample j which is calculated by $e_j = y_j - \hat{y}_j$. Assuming the cost function in feature space is similar to ridge regression, the optimization problem in primal space is written as follows [21]

$$\min_{w,b,e} \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{j=1}^{N} e_j^2 \tag{9}$$

such that $y_j = w^T \varphi(x_j) + b + e_j, j = 1, ..., N$.

It is obvious that if w is infinite, this optimization can not be solved in primal space; thus, the problem is solved in dual space. Assuming $\alpha_j \in \mathbb{R}$ as 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 optimality conditions become as follows

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \to w = \sum_{j=1}^{N} \alpha_j \varphi(x_j) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \to \sum_{j=1}^{N} \alpha_j = 0 \\ \frac{\partial \mathcal{L}}{\partial e_j} = 0 \to \alpha_j = \gamma e_j, j = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_j} = 0 \to y_j = w^T \varphi(x_j) + b + e_j, j = 1, ..., N. \end{cases} \tag{10}$$

After eliminating w and e, the dual problem is obtained as follows

$$\left(\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma} I_N \end{array} \right) \left(\begin{array}{c} b \\ \hline \alpha \end{array} \right) = \left(\begin{array}{c} 0 \\ \hline y \end{array} \right) \tag{11}$$

where Ω is the kernel matrix and Mercer's theorem [22] is applied as follows:

$$\Omega_{jl} = \varphi(x_j)^T \varphi(x_l) = K(x_j, x_l) \quad j, l = 1, 2, \ldots, N. \tag{12}$$

Note that there is no need for explicitly defining the mapping function $\varphi(\cdot)$. This can be done implicitly by positive definite kernel function $K(\cdot, \cdot)$. In this paper, the Radial Basis Function (RBF) is used as a kernel function which is formulated in (1).

Finally, considering α_j and b as the solution for the linear system, the LS-SVM model as a function estimator is obtained as follows

$$\hat{y}(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j) + b. \tag{13}$$

In the case of an RBF kernel, the regularization parameter γ and the kernel parameter σ are tuning parameters. Parameter selection can be done by several methods such as cross-validation and Bayesian learning.

III. Proposed black-box model

A. Data gathering

Weather data are gathered from the weather underground website which is one of the well-known companies in climate and weather forecasting. Data include real measurements for variables like minimum and maximum temperature, precipitation, humidity, wind speed and sea level pressure. These measurements are collected from 70 stations, most of which are located in North America, Europe and East Asia, and cover a time period from the beginning of 2007 until mid 2014. These stations can be seen in Fig. 1.

Furthermore, since this paper aims at forecasting the minimum and maximum temperature form 1 up to 6 days ahead, weather underground predictions of these two variables for these steps ahead are also collected from the website. In the next section, the accuracy of the black-box model predictions for the minimum and maximum temperature are compared with weather underground ones.



Fig. 1: Location of stations used for temperature prediction at Brussels, Belgium.

B. Proposed method

In this section, the proposed algorithm is described. The procedure can be explained in three steps which are depicted in Fig. 2. In the first step, datasets for each city are constructed by assembling the measurements of weather elements for particular dates. The total number of time-series is 70 corresponding to the number of cities, each of them with 27 features and about 2800 observations.

With the aim of localization in mind, some similarity techniques are used to reduce the number of cities. In other words, this reduction ensures that only cities with substantial similarity are favored. In this paper, k-NN based on RBF kernel is used to find the similarities. As a consequence, those cities which have similar time-series to the target city (e.g. Brussels) are selected to be in the model. In this framework, k is a user-defined parameter.

Afterwards, a dataset is generated by concatenating the time-series of the selected cities for the considered time period. In this paper, this dataset is called block $D(t)$, where t is the last day included in the dataset. This block together with the *lag* variable are used as the inputs for time-series box shown in Fig. 2.

Then, having the goal of predicting the future minimum and maximum temperature based on past weather elements included in the dataset, these variables of the target city are forecasted based on a Nonlinear AutoRegressive eXogenous (NARX) [23] model taking into account all past weather elements of all cities. It is obvious that the previous values

of the minimum and maximum temperature of the target city are included in the feature vector. The NARX model is an important type of discrete-time nonlinear system and can be formulated as follows

$$\begin{aligned} y(t+s) \quad &= f(y(t), y(t-1), \ldots, y(t-p), \\ &\quad u(t), u(t-1), \ldots, u(t-q)) + e(t+s) \end{aligned} \tag{14}$$

where $y(t)$ and $u(t)$ are the output and input of the system at time t and $e(t+s)$ is the error of the model at time $t+s$, respectively. The values p and q are the *lag*s, meaning the number of past observations in the time-series that are considered for the prediction task. Note that s denotes the number of steps ahead in the future to predict.

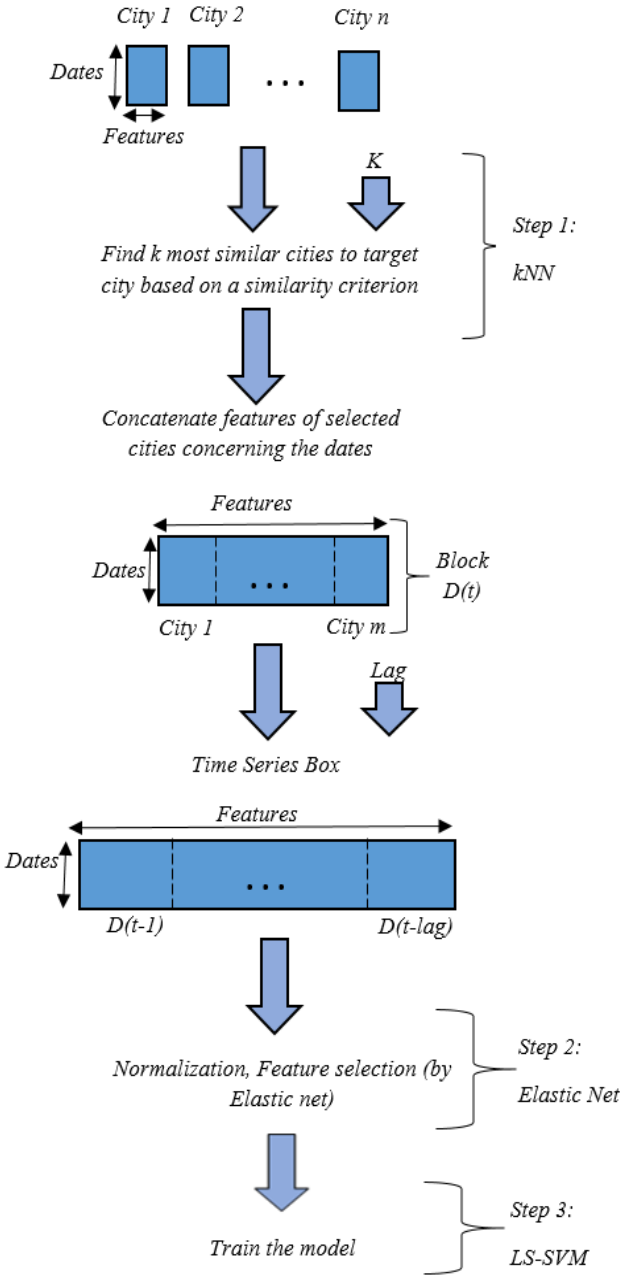

Fig. 2: General framework for the proposed method

Assume that c_{ij} is the jth closest city to city i, $y_i(t+s)$ is the temperature for s steps-ahead and $x_i(t)$ is a vector including all of the features at time t for ith city. Considering $(p=q)$, the weather forecasting model of this paper can be written as following

$$y_i(t+s) \quad = f(x_{\mathcal{S}_i}(t), x_{\mathcal{S}_i}(t-1), \ldots, x_{\mathcal{S}_i}(t-p)) + e(t+s) \tag{15}$$

where $\mathcal{S}_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$ and $i \in \{1, ..., n\}$ and n is the total number of cities.

Note that, although the similarity function is not based on Euclidean distance, the experiments shows the ith city is always in its k-NN set and therefore, the temperature feature of the target city in past steps is already included in $\{x_i(t), \ldots, x_i(t-p)\}$. Therefore, it is considered as a NARX model.

Assume $D(t-p)$ is a $D(t)$ block with p steps delay. As it is shown, a "lag" number of $D(t)$ blocks are integrated to form a larger dataset and prepare it as the input of the feature selection method. The temporal interpretation for $D(t-p)$ is that for a particular day as an observation, the weather elements of p days earlier are considered as feature vector.

It can be seen that after this step, the final number of variables in the feature vector (regressor) is linearly dependent on the lag variable and k and is equal to $lag \times k \times 27$. As it is expected, for small but meaningful values of k and *lag* (e.g. 10) there will be a large number of features. It is undeniable that many of these features might be irrelevant or redundant for the function estimation task. In the proposed method, feature selection is done in step 2. To make data ready for the feature selection procedure, the data are normalized to zero mean and unit variance. In this paper, Elastic net is used as feature selection method. Subsequently, in the last step, a LS-SVM model in trained as function estimator and used to predict target values.

C. Evaluation

In this paper, the evaluation of the accuracy for predictions is based on Mean Absolute Error (MAE) because it is less sensitive to the outliers. Since the values of temperatures are in Celsius, the interpretation of MAE is the average difference between predictions and real values in terms of Celsius degree in the test period. MAE is defined by the following formula

$$MAE = \frac{1}{N_{test}} \sum_{t=1}^{N_{test}} |\hat{y}_i(t) - y_i(t)| \tag{16}$$

where N_{test} is the number of samples (days) in the test set and $\hat{y}_i(t)$ and $y_i(t)$ are predicted and actual values of temperature for ith city at time t respectively.

In addition, the number of days in the test set that the proposed black-box model prediction has an accuracy at least as good as the prediction by weather underground is considered for evaluating the performance of the model. Assume $\hat{y}_i^{BB}(t)$ and $\hat{y}_i^{WU}(t)$ are the predicted values by the black-box model and weather underground. Let N_{test} be the total number of predictions and $N_{\{\hat{y}_i^{BB}(t) \geq \hat{y}_i^{WU}(t)\}}$ be the number of prediction in which proposed method performs at least as good as

weather underground respectively. To do the assessment, Ratio is defined as follows

$$Ratio = \frac{N_{\{\hat{y}_i^{BB}(t) \geq \hat{y}_i^{WU}(t)\}}}{N_{test}} \times 100. \quad (17)$$

Besides, the average values of lag, k in k-NN and ν in Elastic net for 1- to 6-days ahead is discussed in the experiments. Furthermore, in order to analyze the complexity of the models when there is feature selection by k-NN and Elastic net, the percentage of reduction with respect to the total number of features before selection is shown.

IV. Exprimental results

In this section, the results that have been obtained by applying the proposed black-box model on the available data are discussed. First the experimental setup is explained; then the results are presented and discussed.

A. Setup

In this section, the main goal is to compare the accuracy of predictions for minimum and maximum temperature by proposed the black-box model with weather underground forecasts for Brussels. In order to analyze the performance of these methods in various time periods, two different test sets are defined: one from mid-November 2013 until mid-December 2013 (test set Nov/Dec) and the other one from mid-April 2014 to mid-May 2014 (test set Apr/May).

In this paper, k-NN is used to prune and find an effective number of cities. Since the number of advantageous cities is not known in advance, the values of $k \in \{10, 17, 27\}$ are examined.

Furthermore, Elastic net is used as a feature selection method. As shown earlier, for $\nu = 1$ this method represent ridge regression and there is almost no reduction in the number of features, it is considered ν has a value in $[0, 1)$. More specifically, four different values $\{0, 0.2, 0.5, 0.8\}$ are examined for ν. Note that in case of $\nu = 0$, Elastic net is characterizing LASSO.

Considering (15) as the time-series model, the s variable, which is denoting step ahead parameter, is tested for the values from 1 to 6; So, the results are available for multi-days ahead. Also, since the time complexity for checking every possible value for the *lag* parameter p is very high, this parameter is investigated for $p \in \{8, 13, 18, 25\}$ in this paper.

To obtain a good generalization for the models, model selection is done by using 10-fold cross-validation to tune the parameters: the "tunelssvm" function in the LS-SVMlab1.8 is used for tuning γ and σ, the "lasso" function of MATLAB is used for tuning η and grid search is applied for tuning ν, the *lag* variable and k for k-NN.

Note that the model including the training set and its parameter, is updated after one prediction. This means, temporally for the prediction of one particular day temperature, the training set includes all the available real values up to that day. For example, if the temperature for the first of December 2013 is going to be predicted, data from the beginning of 2007 up to end of November 2013 is used to train the model. This training set is also used beyond one day ahead prediction. After one day, when the real values of the features for the first day of December are available on the website, they are added to the training set and a new model is trained to forecast the next day or multi-days ahead temperature.

B. Results

In Fig. 3, four methods are compared regarding to the overall mean absolute error on two test sets together. The performance of weather underground predictions for the minimum and maximum temperature in Brussels is compared with those of the proposed method. In addition, instead of applying the feature selection method on favored cities selected by k-NN, Elastic net reduces the number of features based on all the cities; in other words, step 1 in Fig. 2 is ignored. Moreover, the LS-SVM regression method without feature selection is also tested and compared with the rest (steps 1 and 2 in Fig. 2 are ignored).

As it is depicted, for the minimum temperature, the black-box modeling methods mostly show very good results; among them in many cases the proposed method outperforms. Except in one case (4-days ahead), the proposed method has better prediction accuracy than weather underground. Furthermore, it can be seen that the accuracy of prediction for the maximum temperature is also good, and again the best one is of the proposed method. Nevertheless, for short term prediction (1- to 3-days ahead), weather underground outperforms black-box modeling. Note that the maximum difference between weather underground predictions and the proposed method is 0.75, which occurs in 2-days ahead prediction and implies that the results are still competitive.

In addition, the influence of good feature selection can be seen in comparing the results for LS-SVM alone and Elastic net followed by LS-SVM. In most of the cases feature selection improves the accuracy. The number of selected features highly depends on the parameters η and ν in Elastic net. The larger ν leads to higher number of features to be selected. Experiments show that selected features using Elastic net, is also geographically meaningful. That is, for short term temperature prediction, close cities features are mostly selected, while for long term prediction, the features of further cities are included in model.

The performance of the proposed black-box model predictions for the maximum temperature are not as good as for the minimum temperature. However, the results show a small difference with weather underground prediction in terms of accuracy. Black-box modeling seems to have better performance in long term (more than 3-days ahead) prediction for the maximum temperature.

Tables I and III present the accuracy of methods on each dataset. In most of the cases, the proposed method outperforms weather underground in predicting minimum temperature; this property is more obvious in test set Apr/May in which the worst case happens in 5-days ahead prediction where the accuracies are the same.

Also, it can be concluded that looking at the weather elements locally, instead of globally, which is done by k-NN, improves the performance. In the implementations, it was seen that finding k-NN of the target city regarding to weather elements, is geographically meaningful; e.g. the set of k-NNs of Brussels includes close cities such as Amsterdam,

Berlin, Antwerp while East Asia cities are totally excluded. The selected cities considering different k are shown in Fig. 5.

As it is mentioned before, after each day, the training set is updated and consequently the trained model should be updated as well. In order to have a more accurate model, all of the parameters are tuned and model selection is done again. In Fig. 6 and 7, the average values for k in k-NN, ν in Elastic net and lag variable are plotted for 1- to 6-days ahead prediction in each dataset. It can be observed that for larger days ahead, k is larger and this is more obvious in Apr/May. That is, to have a good prediction of temperature, more cities are needed, e.g. for maximum temperature prediction for 6-days ahead in Apr/May, the best models are generated when the data of approximately 25 closest cities are included in the model, while this number for 1-day ahead is about 11. That is, for long term prediction, the radius of neighborhood is larger. This pattern is geographically meaningful as well because bringing more cities into consideration means looking at further cities features for prediction.

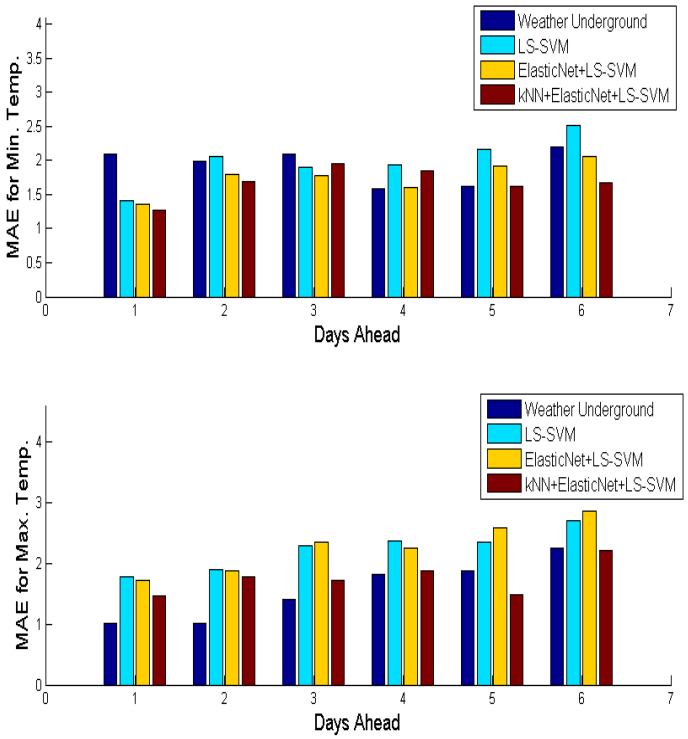

Fig. 3: MAE of the predictions in weather underground, LS-SVM, ElasticNet+LS-SVM and k-NN+ElasticNet+LS-SVM

In addition, it is shown that mostly for larger days ahead, a smaller lag is suitable. So, one may conclude that the relationship between lag and k is inverse. This makes sense because for the larger value of k, more cities are taken into account and as a results $D(t)$ in Fig. 2 has more features and small value of lag can be efficient.

Furthermore, the experiments show that mostly the average ν for predicting minimum temperature is smaller than values of ν for predicting maximum temperatures. Therefore, the Elastic net is closer to LASSO instead of ridge regression and fewer features are selected.

In Tables II and IV, the percentage of days in each dataset that the proposed method performs at least as good as weather underground is shown. Localizing data by using the k-NN method generally increases the number of days that the proposed method beats weather underground. Also, by comparing LS-SVM results and Elastic net followed by LS-SVM, it can be inferred that Elastic net successfully discovers the relevant features. Hence, the number of days that the latter method performs at least as good as weather underground is mostly larger than those of the former.

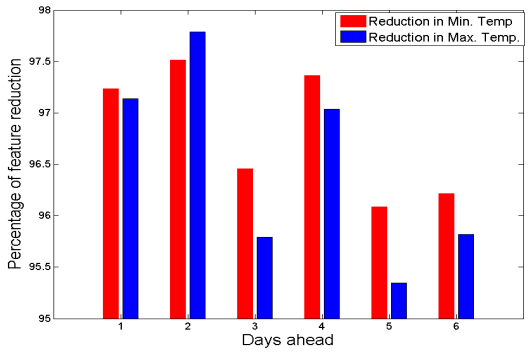

Fig. 4: Average percentage of feature reduction regarding to total number of features.



(a) Cities in the model with k=10



(b) Cities in the model with k=17



(c) Cities in the model with k=27

Fig. 5: Cities in the model with different values of k for k-NN used for temperature prediction in Brussels, Belgium

In Fig. 4 the average percentage of feature reduction for 1- to 6-days ahead prediction is depicted. The bars show how many percent of the total number of variables in feature vector is eliminated from the model. Note that the average number of features before reduction can be computed by $lag \times k \times 27$ using Fig. 6 and 7. As the rate of reduction is more 90%, it can be concluded that the complexity of the model significantly decreases.

Step ahead	Temp.	weather under-ground	All cities (LS-SVM)	All cities (ElasticNet+ LS-SVM)	k-NN+ Elasticnet +LS-SVM
1	Min	1.57	1.38	1.34	**1.15**
	Max	**0.96**	1.35	1.15	1.07
2	Min	**1.57**	1.92	1.76	1.76
	Max	**1.15**	1.69	1.69	1.42
3	Min	**1.76**	1.84	1.92	2.03
	Max	**1.26**	2.15	2.19	1.46
4	Min	**1.23**	1.84	1.65	2.07
	Max	**1.38**	1.92	2.19	1.65
5	Min	**1.76**	1.84	1.92	**1.76**
	Max	1.65	2.15	2.03	**1.19**
6	Min	2.42	2.34	2.30	**1.50**
	Max	2.26	2.03	2.11	**1.73**

TABLE I: MAE of the predictions in weather underground, LS-SVM, Elasticnet+LS-SVM and k-NN+Elasticnet+LS-SVM in test set Nov/Dec.

V. Conclusion

In this paper, a black-box model for weather forecasting was proposed and showed a good performance on predicting minimum and maximum temperature of Brussels in terms of mean absolute error. Also, the predictions of the model were compared with weather underground forecasts and it was seen that they are competitive.

Step ahead	weather under-ground	All cities (LS-SVM)	All cities (ElasticNet+ LS-SVM)	k-NN+ Elasticnet +LS-SVM
1	Min	68%	65%	**73%**
	Max	42%	65%	**73%**
2	Min	50%	**54%**	**54%**
	Max	42%	46 %	**58%**
3	Min	**58%**	54%	**58%**
	Max	42%	42%	**54%**
4	Min	**54%**	42%	46%
	Max	42%	46%	**61%**
5	Min	46%	50%	**54%**
	Max	53%	58%	**80%**
6	Min	50%	62%	**73%**
	Max	58%	62%	**69%**

TABLE II: Ratio for LS-SVM, Elasticnet+LS-SVM and k-NN+Elasticnet+LS-SVM in testset Nov/Dec.

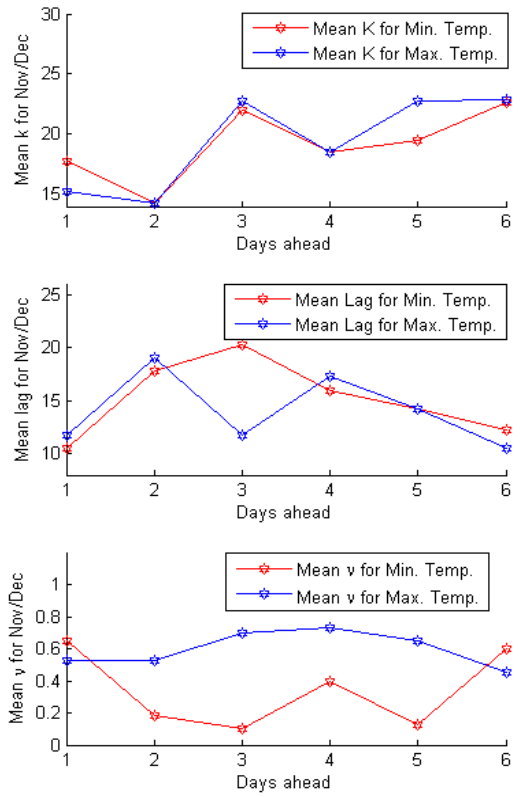

Fig. 6: Average values of k, lag and ν in Nov/Dec for the minimum and maximum temperature.

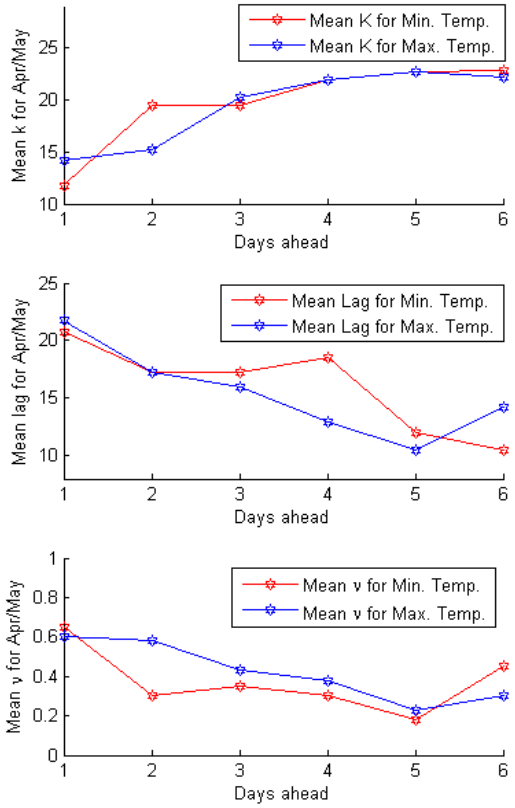

Fig. 7: Average values of k, lag and ν in Apr/May for the minimum and maximum temperature.

Data for a time period of 7 years of 70 cities, mostly from America, Asia and Europe, were gathered from weather underground and in the model k-NN and Elastic net were used to reduce the number of features and decrease the complexity of the model. In addition, LS-SVM was used as a learning method. Results show that each of these components can enhance the accuracy of the model.

Step ahead	Temp.	weather underground	All cities (LS-SVM)	All cities (ElasticNet+ LS-SVM)	k-NN+ Elasticnet +LS-SVM
1	Min	2.59	1.44	1.39	**1.37**
	Max	**1.07**	2.02	2.29	1.85
2	Min	2.37	2.18	1.81	**1.62**
	Max	**0.88**	2.11	2.07	2.11
3	Min	2.40	1.96	**1.62**	1.92
	Max	**1.51**	2.40	2.48	1.96
4	Min	1.92	2.03	1.65	**1.62**
	Max	2.22	2.81	2.29	**2.11**
5	Min	**1.48**	2.25	1.92	**1.48**
	Max	2.07	2.77	3.14	**1.77**
6	Min	2.08	2.66	1.91	**1.85**
	Max	**2.22**	3.37	3.59	2.66

TABLE III: MAE of the predictions in weather underground, LS-SVM, Elasticnet+LS-SVM and k-NN+Elasticnet+LS-SVM in test set Apr/May

Step ahead	weather underground	All cities (LS-SVM)	All cities (ElasticNet+ LS-SVM)	k-NN+ Elasticnet +LS-SVM
1	Min	**78%**	75%	**78%**
	Max	41%	35 %	**51%**
2	Min	77%	81 %	**85%**
	Max	**41%**	40%	**41%**
3	Min	70%	**77%**	70%
	Max	38%	40 %	**51%**
4	Min	63%	77 %	81%
	Max	48 %	48%	**59%**
5	Min	44%	51 %	**100%**
	Max	44%	37%	**70%**
6	Min	40 %	56%	**59%**
	Max	40 %	29%	**55%**

TABLE IV: Ratio for LS-SVM, Elasticnet+LS-SVM and k-NN+Elasticnet+LS-SVM in testset Apr/May.

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). The authors would also like to thank Jeroen de Haas, Gervasio Puertas, Xiaolin Huang, Marco Signoretto, Raghvendra Mall, Vilen Jumutc, Rocco Langone and Ricardo Castro-Garcia for their assistance with this work.

REFERENCES

[1] K. Rasouli, W. W. Hsieh, and A. J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *Journal of Hydrology*, vol. 414, pp. 284–293, 2012.

[2] I. Tasadduq, S. Rehman, and K. Bubshait, "Application of neural networks for the prediction of hourly mean surface temperatures in saudi arabia," *Renewable Energy*, vol. 25, no. 4, pp. 545–554, 2002.

[3] I. Maqsood and A. Abraham, "Weather analysis using ensemble of connectionist learning paradigms," *Applied Soft Computing*, vol. 7, no. 3, pp. 995–1004, 2007.

[4] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. R. Ganguly, "Sparse group lasso: Consistency and climate applications." in *SDM*. SIAM, 2012, pp. 47–58.

[5] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 5, pp. 497–511, 2011.

[6] R. J. Kuligowski and A. P. Barros, "Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks," *Weather and Forecasting*, vol. 13, no. 4, pp. 1194–1204, 1998.

[7] S. Ismail, A. Shabri, and R. Samsudin, "A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10 574–10 578, 2011.

[8] A. Kusiak, H. Zheng, and Z. Song, "Short-term prediction of wind farm power: a data mining approach," *Energy Conversion, IEEE Transactions on*, vol. 24, no. 1, pp. 125–136, 2009.

[9] S. E. Snell, S. Gopal, and R. K. Kaufmann, "Spatial interpolation of surface air temperatures using artificial neural networks: Evaluating their use for downscaling gcms," *Journal of Climate*, vol. 13, no. 5, pp. 886–895, 2000.

[10] N. R. Pal, S. Pal, J. Das, and K. Majumdar, "SOFM-MLP: a hybrid neural network for atmospheric temperature prediction," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 12, pp. 2783–2791, 2003.

[11] M. Signoretto, E. Frandi, Z. Karevan, and J. A. K. Suykens, "High level high performance computing for multitask learning of time-varying models," *IEEE Symposium on Computational Intelligence in Big Data*, 2014.

[12] A. Mellit, A. M. Pavan, and M. Benghanem, "Least squares support vector machine for short-term prediction of meteorological time series," *Theoretical and applied climatology*, vol. 111, no. 1-2, pp. 297–307, 2013.

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

[14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[15] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[16] R. Mall, V. Jumutc, R. Langone, and J. A. K. Suykens, "Representative subsets for big data learning using kNN graphs," *IEEE BigData*, pp. 37–42, 2014.

[17] A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, no. 2, pp. 353–360, 1984.

[18] M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, pp. 65–78, 1982.

[19] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[21] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*. World Scientific, 2002.

[22] J. 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*, pp. 415–446, 1909.

[23] I. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems part i: deterministic non-linear systems," *International journal of control*, vol. 41, no. 2, pp. 303–328, 1985.