Clustering-based feature selection for black-box weather temperature prediction

Zahra Karevan
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
Email : zahra.karevan@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*—Reliable weather forecasting is one of the challenging tasks that deals with a large number of observations and features. In this paper, a data-driven modeling technique is proposed for temperature prediction. To investigate local learning, Soft Kernel Spectral Clustering (SKSC) is used to find similar samples to the test point to be used for training. Due to the high dimensionality, Elastic net is employed as a feature selection approach. Features are selected in each cluster independently and then, Least Squares Support Vector Machines (LS-SVM) regression is used to learn the data. Finally, the predicted values by LS-SVMs are averaged based on the membership of the test point to each cluster. In the experimental results, the performance of the proposed method and "Weather underground" are compared and it is shown that the data-driven technique is competitive with the existing weather temperature prediction sites. For the case study, the prediction of the temperature in Brussels is considered.**

I. Introduction

Accurate weather forecasting is one of the challenges in climate informatics. It involves reliable predictions for weather elements like temperature, 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 utilized for accurate weather prediction and understanding the underlying process. Different types of data-driven methods have been used for weather forecasting both in linear and nonlinear frameworks and among them Artificial Neural Networks (ANN) and Least Squares Support Vector Machines (LS-SVM) are two of the most popular ones. In [2], it is claimed that LS-SVM generally outperforms artificial neural networks. Besides, in our previous works [3], [4], it is shown that LS-SVM performs well for temperature prediction.

Weather forecasting can be considered as a time-series problem which means in order to have an accurate prediction for one particular day, weather variables of some previous days should be taken into account in the prediction model [4]. In this paper, for finding the proper number of previous days that has to be included in the model, Schwarz' Bayesian Information Criterion (BIC) is utilized [5].

Having various weather elements available for several days and locations leads to a large feature vector size and hence, feature selection is essential 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. In this paper, Elastic net, which is a combination of L_1-norm and L_2-norm, is used as the feature selection method. Elastic net establishes a balance between LASSO [6] and ridge regression [7]. Note that if L_2-norm is ignored, Elastic net represents LASSO and if the L_1-norm is disregarded, it corresponds to ridge regression. In this study, Least Squares Support Vector Machines (LS-SVM) [8] are used for modeling. In comparison with SVM, it involves a set of linear equations, instead of convex quadratic programming, to solve the optimization problem.

Mostly, learning methods use all of the data points to train the model. However, local algorithms only use the samples in the area of the test point for model fitting [9]. In this study, the influence of local learning is investigated by finding the similar samples in the training set to the test point prior to the feature selection and learning steps. In order to find a proper sample set for training the model, Soft Kernel Spectral Clustering (SKSC) is used as a clustering approach. SKSC is a fuzzy clustering method based on Kernel Spectral clustering (KSC) [10], but instead of hard clustering, it allows soft membership to the clusters. It uses Average Membership Strength (AMS) criterion for tuning the number of clusters and kernel parameters. Experiments show that SKSC outperforms KSC when the clusters are not well separated.

In this study, the proposed method is used to predict the minimum and maximum temperature in Brussels for 1 to 6 days ahead. Instead of simulated data, the real measurement values of weather elements is used for weather forecasting. In order to avoid missing values, a consistent feature set including real measurements for weather variables such as minimum and maximum temperature, humidity and wind speed is taken into consideration. These features are collected from the weather underground website[1] for 11 stations in the neighborhood of Brussels and cover a time period from the beginning of 2007 until mid 2014.

The remainder of the paper is organized as follows: in the first section, the main components of the proposed method are described. Then, in the second one, these elements are assembled together and the proposed method is explained and finally experimental results are compared with one of

[1] www.weatherunderground.com

the high quality forecasting companies (weather underground) predictions.

II. Background

A. ARMA model and BIC measure

AutoRegressive Moving Average (ARMA) models are widely used in time series problems to estimate a variable based on the linear combination of the previous values. An ARMA model includes two parts [11]: the AR part which shows the number of previous values of the target variable included in the model and MA which shows the previous exogenous variables taken into consideration for function estimation.

Given $y = [y_1, y_2, \ldots, y_N]^T$ 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 and c as a constant, the ARMA model can be written as follows

$$\hat{y}_t = \sum_{j=1}^{p} \zeta_j y_{t-j} + \sum_{h=1}^{q} \nu_h x_{t-h} + x_t + c. \tag{1}$$

Note that the values of p and q are the lag parameters and need to be tuned. Schwarz' Bayesian Information Criterion is one of the popular model selection method which is proposed "for the case of independent, identically distributed observations and linear models" [5]. In this paper, BIC is used to tune the lag (p, q) variable in time series. Hence, it is expressed in the framework of ARMA modeling [12].

Assuming the input distribution belongs to the exponential family, the BIC criterion can be expressed as follows

$$BIC = -2ln(L) + M \times ln(N), \tag{2}$$

where L is the maximized likelihood for the estimated model, N is the number of observations and M is the number of parameters to be estimated. A smaller BIC indicates a better model.

B. Soft Kernel Spectral Clustering

In order to evaluate the performance of using local learning algorithm in weather forecasting application, Soft Kernel Spectral Clustering is utilized to find the similar samples in the training set to the test point. Then, the selected set is used as an input for feature selection and learning modules. SKSC is a fuzzy clustering method with the same core model of Kernel Spectral clustering (KSC) [10], but instead of hard clustering, it allows soft membership to the clusters. It is shown that SKSC outperforms KSC when the clusters are overlapped.

Let k be the number of clusters and $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d \times N}$ where x_i is a vector including d features. Also, consider l is the number of score variables needed to encode the k clusters, $e^{(l)} = [e_1^{(l)}, ..., e_N^{(l)}]^T$ are the projections of the training data in the feature space and $\gamma_l \in \mathbb{R}^+$ is the regularization parameter. $\Phi = [\varphi(x_1)^T, ..., \varphi(x_N)^T]$ is a $N \times d_h$ matrix where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^h$ represents mapping function to a high or infinite dimensional space. Ω is the kernel matrix where $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. Also, $D_\Omega^{-1} \in \mathbb{R}^{N \times N}$ is the inverse of the degree matrix associated to the Ω. The primal formulation of KSC is as follows [10]:

$$\begin{gathered} \min_{w^{(l)}, b_l, e^{(l)}} \tfrac{1}{2} \sum_{l=1}^{k-1} {w^{(l)}}^T w(l) - \tfrac{1}{2N} \sum_{l=1}^{k-1} \gamma_l {e^{(l)}}^T D_\Omega^{-1} e^{(l)} \\ \text{subject to } e^{(l)} = \Phi w^{(l)} + b_l 1_N, l = 1, ..., k-1. \end{gathered} \tag{3}$$

Then, for a given point x_i, the clustering models is as follows

$$e_i^{(l)} = {w^{(l)}}^T \varphi(x_i) + b_l, l = 1, ..., k-1 \tag{4}$$

where b_l is the bias term. The dual problem is formulated as follows

$$D_\Omega^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)} \tag{5}$$

where $\alpha_{(l)}$ is the vector of dual variables, $\lambda_l = \frac{N}{\gamma_l}$, D_Ω is a graph degree diagonal matrix where $d_i^\Omega = \sum_j \Omega_{ij}$ and $M_D = I_N - (1/1_N^T D_\Omega^{-1} 1_N)(1_N 1_N^T D_\Omega^{-1})$ is a centering matrix. For a given data point x_i the dual clustering models is as follow

$$\begin{gathered} e_i^{(l)} = \sum_{j=1}^{N} \alpha_j^{(l)} K(x_j, x_i) + b_l, \\ l = 1, ..., k-1, j = 1, ..., N. \end{gathered} \tag{6}$$

Generally in KSC, there are two types of parameters that have to be tuned: k number of clusters and kernel parameters. In case of Radial Basis Kernel (RBF) (7) the kernel parameter is the bandwidth σ

$$K(x_i, x_j) = \exp(-||x_i - x_j||_2^2 / \sigma^2). \tag{7}$$

Several methods have been proposed for tuning these parameters such as BLF, AMS and modularity [13]. The tuning procedure is based on the grid search and the trained model is evaluated on a separate validation set previously samples from the data. Finally, the combination that yields the maximum criterion is selected. In this study, Balanced Line Fit (BLF) and Average Membership Strength (AMS) are used for model selection.

In the Balanced Line Fit method, the collinearity in the space of the projections between the validation and the training samples that are in the same cluster is computed. The maximum value for BLF criterion is 1 and is achieved when the clusters are well separated. Note that the higher BLF indicates better clustering.

$$BLF(D^V, k) = \mu\, linefit(D^V, k) + (1 - \mu)\, balance(D^V, k). \tag{8}$$

In (8) D^V is the sampled validation set and k is the number of clusters. $\mu \in [0, 1]$ is a parameter giving weights to the $linefit$ and $balance$. The $linefit$ index is 0 when the distribution of the score variable is spherical and equals to 1 when the score variables are collinear. On the other hand, the $balance$ index tends to be 1 when the clusters have equal number of samples and is 0 when they don't have the same number of points. One of the drawbacks of this method is that there is no specific way to select μ value. Moreover, when the clusters are overlapped, the assumption of having linear structure in projection space can not be hold any more.

SKSC leverages KSC in the initialization which means in the first step, SKSC uses KSC to identify the first clusters in

the data and then improves the clusters by re-calculating the prototypes in the score variable space. Finally, each sample is assigned to a cluster based on its distance with the prototype. To avoid the drawbacks of BLF, SKSC utilizes Average Membership Strength for model selection. In AMS, the mean membership value for the validation points to each cluster is calculated. Note that the membership degree shows the certainty with which a sample belongs to the clusters. To find the membership value for each sample, the cosine similarities between the data point and the prototypes of the clusters are computed. For a given data point x_i, the membership value to the cluster m is as follows [14]

$$cm_i^{(m)} = \frac{\prod_{j \neq m} d_{ij}^{cos}}{\sum_{h=1}^{k} \prod_{j \neq h} d_{ij}^{cos}}, \sum_{h=1}^{k} cm_i^{(m)} = 1, \tag{9}$$

where k is the number of cluster and d_{ij}^{cos} is the cosine similarity between sample ith and the prototype of the cluster j in score variables space.

$$AMS = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{N_j} \sum_{i=1}^{N_j} cm_i^{(j)} \tag{10}$$

where N_j is the number of samples in cluster j.

C. Elastic net

To deal with the high dimensionality of the dataset, the feature selection becomes an essential step for obtaining the relevant features. Two of the most popular methods for reducing the number of features are Elastic net [15] and LASSO [6]. Considering x as the feature vector and $x_{(i)}$ be the ith feature, the linear regression model is expressed as follows

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{(1)} + \ldots + \hat{\beta}_d x_{(d)} \,. \tag{11}$$

Several methods have been proposed for estimation of $\hat{\beta}$ 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. In addition to continuous shrinkage, LASSO attempts to produce sparse model and is being used as a feature selection method. In comparison with Ordinary Least Squares (OLS), the sparse model can provide a better interpretation for the embedded system. Furthermore, it may improve the prediction accuracy by increasing the bias and reducing the variance of the predicted values. Nevertheless, it has its own limitations. For example, in the case of highly correlated features, LASSO chooses only one of them, no matter which one. Moreover, if the number of samples is smaller than number of features, 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 vectors including d features and the response value at observation i respectively. Elastic net solves

$$\hat{\beta} = \arg\min_{\beta} J(\beta, \lambda_1, \lambda_2), \tag{12}$$

where

$$J(\beta, \lambda_1, \lambda_2) = ||y - X^T\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||^2, \tag{13}$$

with

$$||\beta||^2 = \sum_{j=1}^{p} \beta_j^2, ||\beta||_1 = \sum_{j=1}^{p} |\beta_j|. \tag{14}$$

In equation (13), λ_1 and λ_2 are penalty parameters. Let $\nu = \lambda_2/(\lambda_2+\lambda_1)$ then the Elastic net minimization is an equivalent form of

$$\hat{\beta} = \arg\min_{\beta} ||y - X^T\beta||^2, \tag{15}$$

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

The term $(1-\nu)||\beta||_1 + \nu||\beta||^2$ is the Elastic net penalty and is a convex combination of L_1-norm and L_2-norm. Considering $\nu = 1$, the optimization formula becomes ridge regression [7], while for $\nu = 0$, it represents LASSO. In this paper, it is assumed that $\nu \in [0, 1)$.

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

D. Least Squares Support Vector Machines

In this paper, Least Squares Support Vector Machines (LS-SVMs), proposed in [16] [8], are used to learn the data. In comparison with quadratic programming in Support Vector Machines, LS-SVM 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:

$$y(x) = w^T\varphi(x) + b \tag{16}$$

where $b \in \mathbb{R}$ and the dimension of w depends on the feature map and is equal to h. The optimization problem in primal space is written as follows [8]

$$\begin{aligned} &\min_{w,b,e} \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{j=1}^{N} e_j^2 \\ &\text{subject to } y_j = w^T\varphi(x_j) + b + e_j, j = 1, ..., N, \end{aligned} \tag{17}$$

where $\{x_j, y_j\}_{j=1}^N$ is the training set, γ is regularization parameter and $e_j = y_j - \hat{y}_j$ is the error between the actual and predicted output for sample j.

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 are expressed as follows

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{j=1}^{N} \alpha_j \varphi(x_j) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{j=1}^{N} \alpha_j = 0 \\ \frac{\partial \mathcal{L}}{\partial e_j} = 0 \rightarrow \alpha_j = \gamma e_j, j = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_j} = 0 \rightarrow y_j = w^T \varphi(x_j) + b + e_j, j = 1, ..., N. \end{cases} \tag{18}$$

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

$$\left(\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma} I_N \end{array} \right) \left(\frac{b}{\alpha} \right) = \left(\frac{0}{y} \right) \tag{19}$$

where Ω is the kernel matrix and Mercer's theorem [17] is applied as follows:

$$\Omega_{jl} = \varphi(x_j)^T \varphi(x_l) = K(x_j, x_l) \quad j, l = 1, 2, \ldots, N. \tag{20}$$

Note that there is no need for explicitly defining the mapping function $\varphi(\cdot)$. This can be done implicitly by positive definite kernel function $K(\cdot, \cdot)$. There are different type of functions which can generate kernel matrix. In this paper, the Radial Basis Function (RBF) is used as a kernel function which is formulated in (7). In this case, the regularization parameter γ and the kernel parameter σ are tuning parameters.

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

$$\hat{y}(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j) + b. \tag{21}$$

III. Clustering based feature selection

A. Data gathering

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, precipitation, humidity and pressure from the beginning of 2007 until mid 2014 and for 11 cities including Brussels, Liege, Antwerp, Amsterdam, Eindhoven, Dortmund, London, Frankfurt, Groningen, Dublin and Paris.

Moreover, 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 in the test period are also collected from the website. In the experiments part, the performance of the proposed method is also compared with accuracy of the weather underground company in temperature prediction. The number of samples is equal to the number of days from the beginning of 2007 until the last day for each the real measurement for weather elements is available. Also, there are 18 measured weather variables for each day in each location.

B. Proposed method

In this section, the methods explained in the background are merged together to form a data-driven modeling for weather temperature prediction. With the aim of predicting the future minimum and maximum temperature, these values can be forecasted based on past weather variables included in the dataset. It is obvious that the previous values of the minimum and maximum temperature of the target city are included in the feature vector. The model can be written as follows

$$\begin{aligned} \hat{y}_{t+s} = f(&y_t, y_{t-1}, \ldots, y_{t-p}, \\ &x_t, x_{t-1}, \ldots, x_{t-q}) \end{aligned} \tag{22}$$

where y_t and x_t are the output and input of the system at time t and s is positive integer denoting the number of steps ahead in the future to predict, respectively. The value q and p are the *lag* parameters, indicating the number of past observations and system output in the time-series that are considered for the prediction task. Consequently, the feature vector includes all of the collected features from all of the stations for a particular day. Thus, dataset is generated by concatenating the time-series of the locations for the considered time period and is shown in Fig. 1 by block $D(t)$, where t is the last day included in the dataset. Note that the output of the system y_t is the temperature variable in Brussels and is included in the feature vector.

It is obvious that $D(t - lag)$ is a $D(t)$ block with lag steps delay. As it is shown, a "lag" number of $D(t)$ blocks are integrated to form a larger dataset $X^{new} = [x_1^{new}, x_2^{new}, \ldots, x_N^{new}] \in \mathbb{R}^{d' \times N}$ which is used as the input of the feature selection method. The total number of features d' in this case equals to $lag\times$ (number of stations)$\times$ (number of features in each station). Note that the output of the system can be written as

$$\hat{y}_{t+s} = f(x_t^{new}). \tag{23}$$

In the proposed method, Elastic net is used as a feature selection. As previously mentioned, Elastic net fulfills feature selection task by fitting a linear model. This is the motivation to look into the $f(\cdot)$ function in (22) as a linear model. The linear formulation can be expressed as follows

$$\hat{y}_{t+s} = \sum_{j=1}^{p} \zeta_j y_{t-j} + \sum_{r=1}^{q} \nu_r x_{t-r} + c, \tag{24}$$

where c is a constant value. In this paper, for the simplicity p is considered to be equal to q. As it can be noticed, the general structure of the model is similar to ARMA model (1); thus, BIC is employed for tuning q in similar strategy that is used for ARMA.

Mostly, learning methods looks globally into the data which means they use all of the samples for model fitting. Seasonal behavior of the temperature is an intuitive reason to investigate local learning algorithms. In local learning, instead of using all of the samples for training the model, only those who are in the region of the test point are used for model fitting. In this study, the main steps are similar to [9]: First, for each test point, similar training samples using SKSC are selected.

Then, these samples are used as an input for feature selection module.

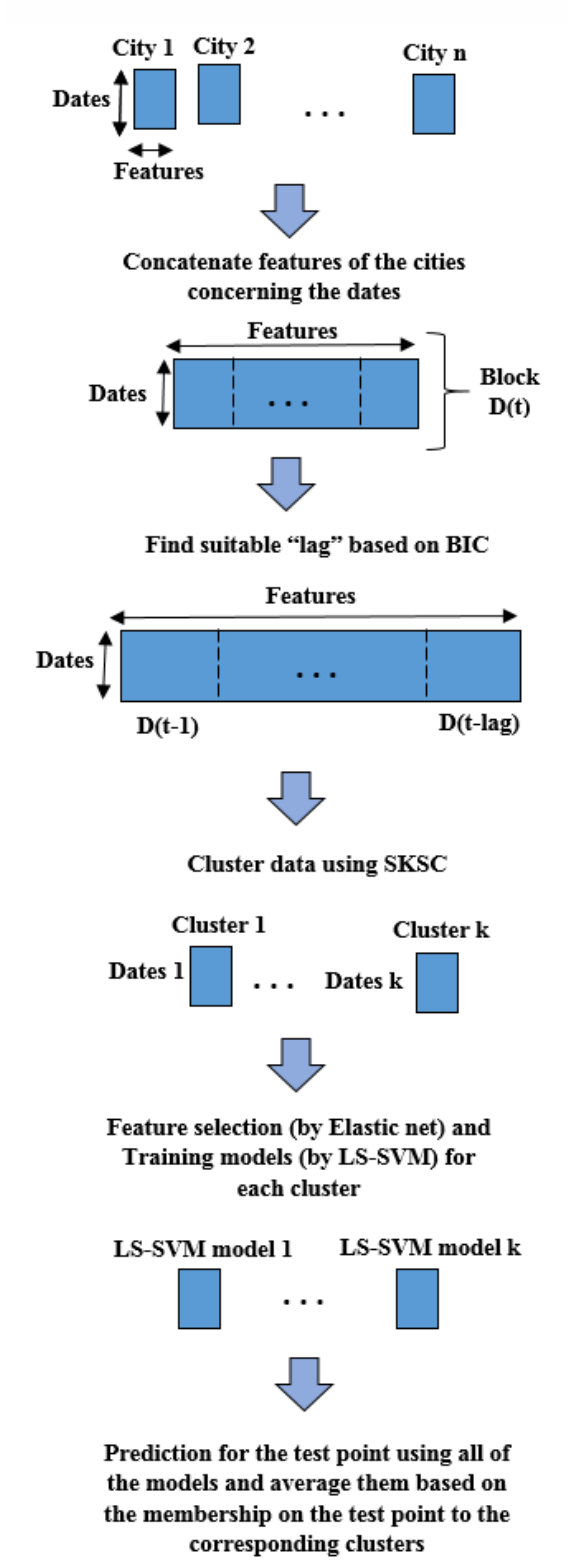


Fig. 1: General scheme of the proposed method.

Since soft clustering is used for sampling, for each test sample there is a membership value assigned to each cluster. This may give the opportunity to use all of the data points to have a good prediction. Assume that training samples have different effects on the prediction task based on their similarity to the test sample. Therefore, different weights are given to each cluster based on the test membership values. For the samples in each cluster, the feature selection procedure is done independently and then different LS-SVM models are trained. Afterwards, the prediction for the test point is done by all of the LS-SVM models, and finally based on the corresponding membership values to the clusters, the weighted average of the prediction is computed as follows

$$\hat{y}_{t+s} = \sum_{m=1}^{k} \hat{y}_{t+s}^{(m)} \times cm_t^{(m)}, \tag{25}$$

where

$$\hat{y}_{t+s}^{(m)} = f(x_t^{new(m)}). \tag{26}$$

In (26), $x_t^{new(m)} \in \mathbb{R}^{d'_m}$ where d'_m is the number of selected features in cluster m. Note that, $cm_t^{(m)}$ is the membership value of the test point x_t^{new} to the corresponding clusters which can be found by equation (10), and the function $f(\cdot)$ is estimated by LS-SVM.

IV. Experiments

In this section, the performance of the proposed method for minimum and maximum temperature forecasting is compared with weather underground predictions for Brussels. Same as our previous work [4], in order to evaluate the performance of the data-driven methods in different time periods, two independent 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).

There are some parameters that have to be tuned: in Elastic net the variables ν, which balanced between $L1$-norm and $L2$-norm, and η, in the constrain condition, are tuned by cross-validation. In addition, the LS-SVM parameters which include the kernel bandwidth σ and the regularization parameter γ are also tuned by 10-fold cross-validation using "tunelssvm"' function in LS-SVMlab1.8 toolbox.

To exploit all of the available data, after each day, the training set is updated and as a result the trained model should be updated as well. In order to have a better performance, all of the parameters should be tuned again. Due to the time complexity, in this paper, the updating is done on a weekly basis.

A. Evaluation

Same as our previous work [4], due to less sensitivity to the outliers, Mean Absolute Error (MAE) is used for the evaluation of the performance. Note that the values of temperatures are in Celsius, MAE denotes the average difference between predictions and real values in terms of Celsius degree in the test period. MAE is defined by the following formula

$$MAE = \frac{1}{N_{test}} \sum_{t=1}^{N_{test}} |\hat{y}_t - y_t| \tag{27}$$

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

The comparison between the performance of KSC and SKSC is based on the Silhouette criterion which compares the similarity of each data point to other samples in its own cluster with the similarity to the samples in other clusters. Considering d_i^{same} to be the average distance of the given sample x_i to the samples in its own cluster and d_i^{diff} be the average distance of x_i to samples in other clusters, the Silhouette value can be calculated as follows

$$S_i = \frac{d_i^{same} - d_i^{diff}}{max(d_i^{same} - d_i^{diff})}. \quad (28)$$

For the Silhouette criterion, the higher value shows better clustering solution.

B. Results

In Tables I and II, the MAE of four methods are compared in both test sets. As it is shown, the performance of weather underground (WU) predictions for the minimum and maximum temperature in Brussels is compared with the following scenarios: first, "LS-SVM" is used to learning the data with all of the features, then "ENet + LS-SVM" where Elastic net is used as a feature selection method in a global learning scenario and then LS-SVM is used for learning, and finally "Clu + ENet + LS-SVM" which is the proposed method.

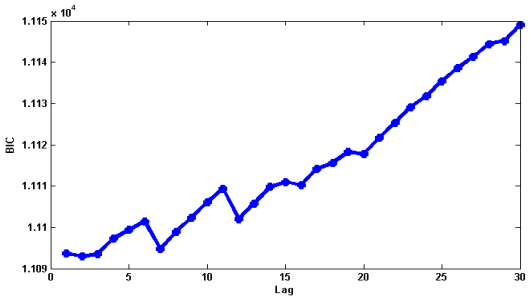
It can be concluded that for the minimum temperature, the data-driven approaches mostly outperform weather underground company. In the case of maximum temperature, the performance is not as good as the minimum temperature prediction, but it is still competitive with the one of weather underground. In particular, the influence of localizing the data can be seen by comparing the results for two last columns. For both minimum and maximum temperature prediction, among the data-driven methods, the best performance is mostly observed for the proposed method case. This means that with the help of the clustering, better features can be selected.

Step ahead	Temp.	WU	LS-SVM	ENet+LS-SVM	Clu + ENet + LSSVM
1	Min	1.57	1.57	1.43	**1.26**
	Max	**0.96**	1.35	1.29	1.19
2	Min	**1.57**	**1.57**	1.69	1.69
	Max	**1.15**	1.46	1.57	1.46
3	Min	1.76	**1.61**	1.81	1.88
	Max	**1.26**	1.65	1.69	1.73
4	Min	**1.23**	1.84	1.79	1.84
	Max	**1.38**	2.07	1.88	1.92
5	Min	**1.76**	1.92	**1.76**	1.88
	Max	1.65	1.88	1.69	**1.46**
6	Min	2.42	2.34	**2.18**	2.21
	Max	2.26	1.88	1.76	**1.61**

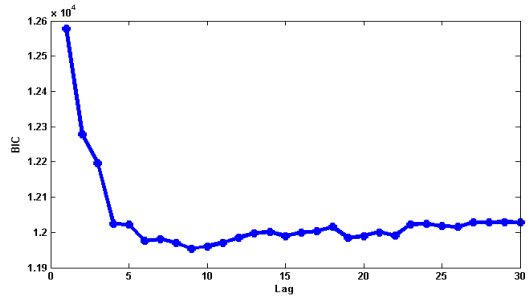
TABLE I: MAE of the predictions in Weather Underground(WU), LS-SVM, Elasticnet+LS-SVM and SKSC+Elasticnet+LS-SVM in test set Nov/Dec.

Step ahead	Temp.	WU	LS-SVM	ENet+LS-SVM	Clu + ENet + LSSVM
1	Min	2.59	1.46	1.51	**1.36**
	Max	**1.07**	2.22	2.18	2.07
2	Min	2.37	2.15	1.92	**1.76**
	Max	**0.88**	2.29	2.29	2.18
3	Min	2.40	2.03	2.03	**1.88**
	Max	**1.51**	2.37	2.57	2.37
4	Min	**1.92**	1.96	2.07	**1.92**
	Max	2.22	2.40	2.36	**2.18**
5	Min	**1.48**	2.18	2.29	2.03
	Max	**2.07**	2.51	2.57	2.14
6	Min	2.08	2.33	2.18	**2.03**
	Max	2.22	2.40	2.49	**2.11**

TABLE II: MAE of the predictions in Weather Underground(WU), LS-SVM, Elasticnet+LS-SVM and SKSC+Elasticnet+LS-SVM in test set Apr/May.



(a) BIC for 1 day ahead prediction



(b) BIC for 6 day ahead prediction

Fig. 2: BIC for identically different optimal lag values for 1 and 6 days ahead prediction.

In Fig. 2 the BIC values for different lag values for 1 and 6 days ahead are shown. Obviously, for long term prediction the larger lag value gives better performance. Moreover, it seems that the different values of this parameter are good candidates to be chosen. Hence, the performance of the proposed method is evaluated for different lag values. Defining the lag value in the range of 7 to 20 seems to be a reasonable.

Fig. 3 shows the AMS values for different number of clusters when SKSC is applied. Evidently, smaller number of clusters provides better clustering. In all of the cases, the maximum AMS is achieved when the number of clusters is 2. In this case, as it is shown in Fig. 4, the clusters can be remarked as summer and winter ones.

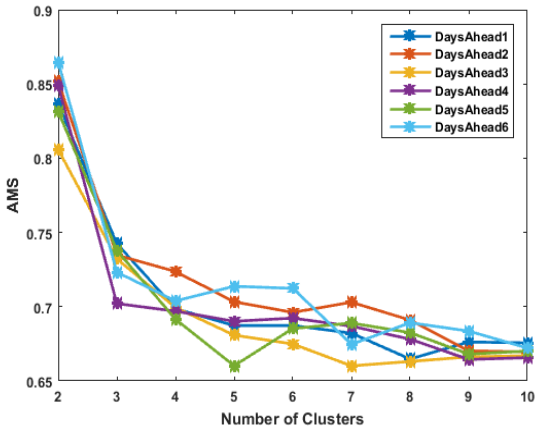
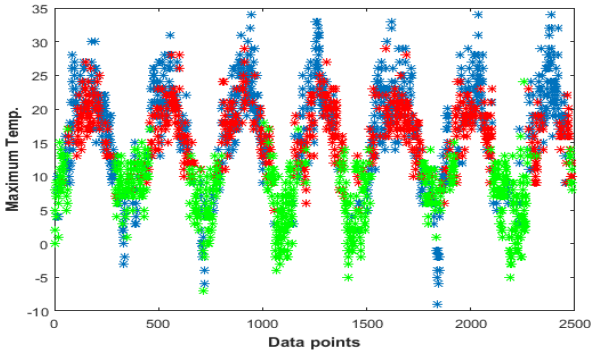
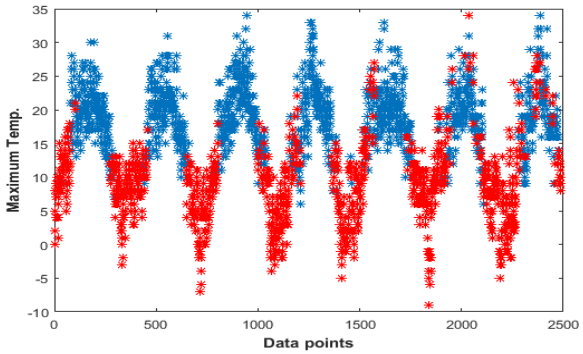


Fig. 3: AMS value for different number of clusters.



(a) Clustering using KSC

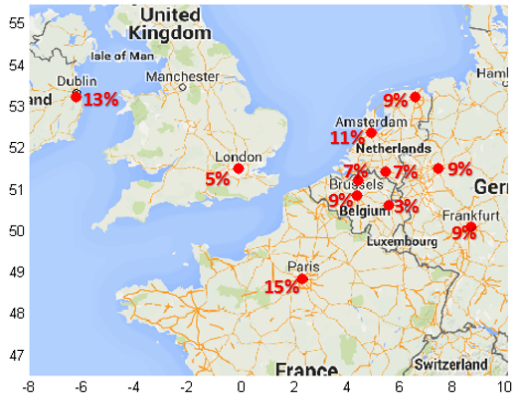


(b) Clustering using SKSC

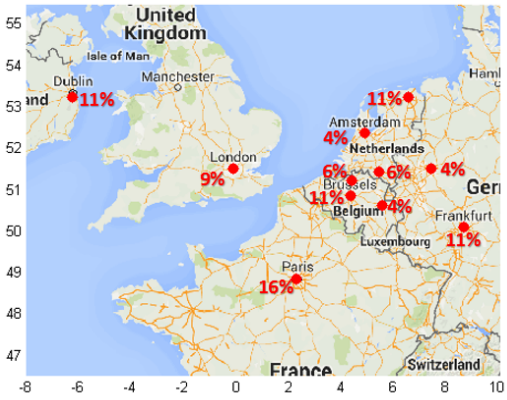
Fig. 4: Comparison between KSC and SKSC.

Step Ahead	1	2	3	4	5	6
Silhouette KSC	0.12	0.9	0.11	0.10	0.12	0.8
Silhouette SKSC	0.36	0.29	0.31	0.29	0.28	0.33

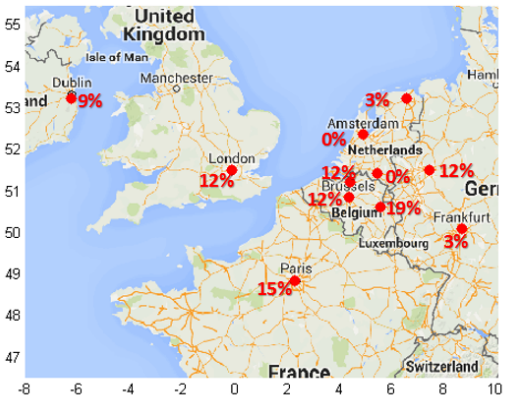
TABLE III: Comparison for Silhouette values for 1 to 6 days ahead predictions.



(a) The percentage of selected features using all of the samples



(b) The percentage of selected features per city in winter cluster



(c) The percentage of selected features per city in summer cluster

Fig. 5: Comparison between the percentage of the number of selected features per city in global (a) and localization with clustering (b,c) scenarios.

In Fig. 4, the maximum temperature of the samples based on their clusters using KSC and SKSC is depicted. KSC identifies three embedded clusters, while SKSC finds two which are winter and summer clusters. In Table III, the Silhouette criterion of the clustering results using KSC and SKSC for different step ahead prediction is shown. In all of

the cases SKSC outperforms KSC. Thus, it can be concluded that the models selection based on AMS is more efficient that BLF.

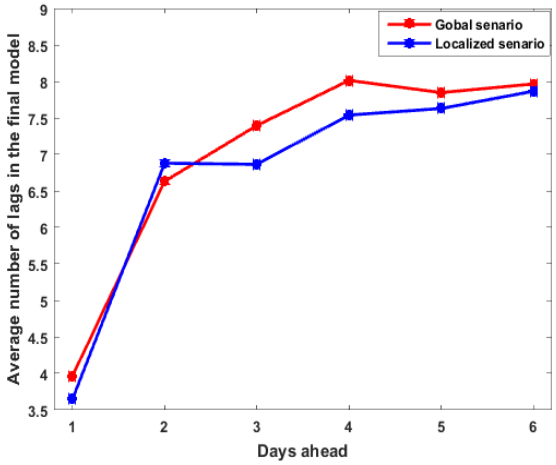


Fig. 6: Average number of lags in final model for both global and localized scenarios

Fig. 5 is an example for LASSO case and shows the percentage of the number of selected features in each city with respect to the total number of selected features. It can be seen that looking into the data in the seasonal (clustered) way can cause different features to be selected. As is depicted, the possible different impacts of the cities on Brussels in different time periods are considered in the seasonal scenario and this phenomena can improve the forecasting performance.

In Fig. 6, the average number of lags from which features are selected for the maximum temperature prediction is shown. It is obvious that both in global or localized scenarios, for long-term prediction, a larger lag is required for accurate forecasting. The pattern is the same for the minimum temperature prediction.

V. Conclusion

In this paper, a data-driven modeling technique is proposed for temperature prediction. To exploit the advantages of the local learning, Soft Kernel Spectral Clustering (SKSC) is utilized to find similar samples to the test point to be used as the training set. Experiments show that SKSC gives better performance than KSC and partitions the data in two clusters corresponding to the winter and summer seasons. Feature selection and learning the data are done independently in each cluster and the results are combined based on the membership value of the test point to the corresponding clusters. For the case study, the prediction of the minimum and maximum temperature in Brussels is considered. Experiments show that the performance of the proposed method is comparative with the predictions of weather underground company.

Acknowledgments

EU: 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] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[2] 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.

[3] 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.

[4] Z. Karevan, S. Mehrkanoon, and J. A. K. Suykens, "Black-box modeling for temperature prediction in weather forecasting," in *International Joint Conference on Neural Networks*, 2015, pp. 1–8.

[5] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

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

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

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

[9] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural computation*, vol. 4, no. 6, pp. 888–900, 1992.

[10] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transactionson Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, 2010.

[11] A. Pankratz, *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons, 2009, vol. 224.

[12] E. P. Clement, "Using normalized bayesian information criterion (BIC) to improve box-jenkins model building," *American Journal of Mathematics and Statistics*, vol. 4, no. 5, pp. 214–221, 2014.

[13] R. Langone, R. Mall, C. Alzate, and J. A. K. Suykens, "Kernel spectral clustering and applications," in *Unsupervised Learning Algorithms*, M. E. Celebi and K. Aydin, Eds. Springer International Publishing, 2016 (in press).

[14] R. Langone, R. Mall, and J. A. K. Suykens, "Soft kernel spectral clustering," in *International Joint Conference on Neural Networks*, 2013, pp. 1–8.

[15] 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.

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

[17] 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.