Combining environmental information from multiple and disparate data sources for minimum temperature prediction

Sergio Hernández, Sergio Baltierra and Guillermo Becerra
Laboratorio de Procesamiento de Información Geoespacial.
Universidad Católica del Maule.
Talca, Chile.
Email: shernandez@ucm.cl

Philip Sallis
Geoinformatics Research Centre.
Auckland University of Technology.
Auckland, New Zealand.
philip.sallis@aut.ac.nz

***Abstract*—Minimum temperature forecasts are a highly valuable tool in agricultural decision making. In this paper, a method for statistical forecasting of minimum temperature predictions using multiple data sources is given. The proposed approach uses a well known method for non-linear regression such as Gaussian Process (GP) and produces a prediction for the minimum temperature for the next day. In addition to the forecast at a single spatial location, a mixture of GP (MGP) models is proposed in order to combine data from multiple independent data sources. The method is tested with real data collected from wireless sensor networks and preliminary results shows improved performance when using the MGP approach while also allowing a fast E-M algorithm for parameter estimation using the product of several predictive distributions.**

Keywords*—*frost prediction; Gaussian Processes; Mixture Models; Distributed Machine Learning.

I. Introduction

Agroclimatological monitoring and prediction is an important tool for agricultural producers in order to anticipate and mitigate extreme climatological events. Understanding the occurrence of such events is essential in order to protect crops, especially during the growing season when faced with extreme cold such as frosts. Recent advances in sensors networks enables the estimation of the trend in minimum temperatures, so when they reach a critical level, the producer can protect crops ahead [12].

Currently, there are several empirical studies that attempt to predict or anticipate the occurrence of the minimum temperature at a given time. One of these is the study by the United Nations Organization for Food and Agriculture (FAO) [10] shows a linear regression example implemented in a spreadsheet, where a producer can track the occurrence of the minimum temperature in plain areas. These methods deliver descriptive models, however calibration is required and the empirical coefficients have to be estimated to account for the time of the year and the local conditions.

Other works address this problem with models based on neural networks, for example in [3], [4], the authors describe a method to predict the occurrence of spring frosts in agricultural mountainous areas from meteorological data such as temperature, relative humidity, solar radiation, wind direction and speed. The authors reported and a the requirement of a different calibration for each station.

More recently, an ensemble neural network was used to achieve temperature predictions throughout an entire year with improved mean absolute error when compared to specific models [9]. In [2], a web based fuzzy expert system is used to develop frost warnings from observed meteorological conditions and expert knowledge is used to develop fuzzy logic rules for different scenarios.

In this paper we predict minimum temperature for the next day using multiple and disparate sources of environmental data. This problem requires not only to be able to analyze a single source of data and being able to perform a prediction at the same location, but also being able to gather data from multiple and geographically distributed locations and generalize to test locations.

II. Combining environmental information

Minimum temperature prediction models require a vast amount of training data in order to produce reliable results. For example, a method for year-round temperature prediction with artificial neural networks used 1.25 million training patterns [9]. SVR is another approach for prediction that is able to generalize from smaller training data sets. The SVR framework performs structural risk minimization and is therefore able to minimize empirical error and model complexity and thus avoiding overfitting. This model was reported as having improved accuracy when compared to the artificial neural network model while also requiring less data for training [1].

Gaussian Process (GP) priors can be used as an alternative to SVR for regression and classification models [6]. Rather than minimizing the structural risk of a regression model for a set of input $\mathbf{X}$ and output data $\mathbf{y}$, the GP approach defines a prior $p(f|\mathbf{X},\mathbf{y})$ over the unknown model $f(x)$ and therefore can be seen as a Bayesian alternative to SVR. Once we have new test data $\mathbf{x}_*$, the posterior distribution of y_* can be computed by marginalization:

$$p(y_*|\mathbf{x}_*,\mathbf{X},\mathbf{y}) = \int p(y_*|f,\mathbf{x}_*,\mathbf{X},\mathbf{y})p(f|\mathbf{X},\mathbf{y})df \quad (1)$$

The GP prior can be expressed as $f(x) \sim GP(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$, where $m(x) = \mathbb{E}[f(\mathbf{x})]$ is a mean function and $K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ is the kernel or covariance function. In the case of noisy regression problems such as $y = f(\mathbf{x}) + \epsilon$ with Gaussian additive noise $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$, the covariance between any two training points can be written as:

$$\begin{aligned} cov(\mathbf{x}_i, \mathbf{x}_j) &= K(\mathbf{x}_i, \mathbf{x}_j) + \sigma_y^2 \delta_{ij} \\ &\equiv K_y \end{aligned}$$

Assuming $m(x) = 0$ and a single test point $\mathbf{x}_*$, we can use the GP prior to estimate the posterior distribution of $f_* = f(\mathbf{x}_*)$ as:

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_*|\mathbf{k}_*^T K_y^{-1}\mathbf{y}, k_{**} - \mathbf{k}_*^T K_y^{-1}\mathbf{k}_*) \quad (2)$$

The performance of the GP method largely depends on the choice of the kernel and the kernel parameters or hyper-parameters. There is a variety of covariance functions available, leading to functions with different degrees of smoothness. One typical covariance function is the so called squared-exponential kernel:

$$k_y(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(-\frac{1}{2}\sum_d \nu_d^2 (x_{id} - x_{jd})^2) + \sigma_y^2 \delta_{ij} \quad (3)$$

The hyper-parameters $\theta = (\sigma_f, \sigma_y, \nu_1, \ldots, \nu_d)$ can be obtained by resorting to Markov Chain Monte Carlo (MCMC) techniques or directly maximizing the marginal data log-likelihood:

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}\mathbf{y}^t K_y \mathbf{y} - \frac{1}{2}\log|K_y| - \frac{N}{2}\log 2\pi \quad (4)$$

It is important to notice that GP priors using this kernel are not appropriate when the function has discontinuities or the degree of smoothness is not stationary. In our case, meteorological data comes from multiple and geographically distributed sources so data is horizontally partitioned.

Another problem with the GP methodology is the computational complexity which is $O(N^3)$ for N training patterns. Compared to winter-only models, year-round minimum temperature predictions require a massive amount of data, making GPs not well suited for the prediction problem.

However, if only have access to a small portion of the data, so apart from modeling heterogeneity we would also like to model discontinuities during the year.

A. Distributed Learning using Mixtures of Gaussian Processes

The basic GP methodology assumes a single dataset and a single function $f(\cdot)$ with global hyper-parameters used to generate the output patterns. Mixture models and mixture of experts have been proposed for non-stationary data [5]. These models are able to learn from hetereogenous data sources by modeling the joint distribution of multiple regression or classification models.

Mixture of GPs (MGP) have been proposed as an alternative to the mixture of experts model for dealing with non-stationary data [11]. The idea behind the MGP approach is to use local GP models on different subsets of the input data and a gating network that is responsibile of assigning input data points to one of the M local experts.

Several gating networks, treatment of hyper-parameters and inference procedures for the MGP model have been proposed in the literature. For example. the gating network used in the infinite MGP [7], [14] requires to estimate parameters that scale linearly with the dimension of the inputs (length-scale parameters) and whose inference is only available with MCMC or variational techniques.

An alternative formulation for the gating network of the MGP model can be found in [8]. In this case, we consider a set $\mathcal{S} = \{S_1, \ldots, S_M\}$ models which are responsibile of entire curves and each curve represents a single batch of measurements. In this case, the gating network is entirely based on the joint distribution of the input $\mathbf{X}$ and the output space $\mathbf{y}$. The GP prior for the $m-th$ batch of data can be written as follows:

$$f_m(x) \sim \sum_i^M \pi_i \, GP(m_i(\mathbf{x}), K_i(\mathbf{x}, \mathbf{x}')) \quad (5)$$

The parameters of the MGP model now include a set of hyper-parameters for each model $\mathbf{\Theta} = (\theta_1, \ldots, \theta_M)$ and the gating network $\pi = (\pi_1, \ldots, \pi_M)$. A Bayesian approach for hyper-parameters inference was adopted in [8], however an efficient procedure based on the Expectation-Maximization (EM) algorithm was further proposed in [13].

B. E-M for Mixtures of Gaussian Processes

The E-M algorithm is an iterative procedure for handling models with missing variables. In this case the missing data is a set of binary variables $\mathbf{z} = \{z_1, \ldots, z_m\}$ that indicate that the data point was generated from the $m-th$ batch, and whose expectation $\mathbb{E}[z_m = \mathbf{1}] = \gamma(\boldsymbol{z})$ is used to maximize the data likelihood:

$$p(\mathbf{X}, \mathbf{y}|\mathbf{\Theta}, \pi) = \prod_n \sum_m \pi_m p(y_n|\mathbf{x_n}, \theta_\mathbf{m}) \quad (6)$$

There is no closed form for the indicator parameters $\boldsymbol{z}$, however the approach taken in the E-M algorithm is to calculate the expectation:

$$\gamma(\boldsymbol{z}_n) \propto \pi_m p(y_n|\mathbf{x}_n, \theta_m) \quad (7)$$

Using this expectation, the parameters π can be derived as:

$$\pi_m = \frac{1}{N}\sum_n \gamma(\boldsymbol{z}_n) \quad (8)$$

The set of GP hyper-paraneters $\mathbf{\Theta}$ can now be obtained by differentiating $\frac{\log p(\mathbf{y}_c|\mathbf{X}_c,\theta_m)}{\theta_m}$, where $\mathbf{y}_c = \gamma(\boldsymbol{z})\mathbf{y}$ and $\mathbf{X}_c = \gamma(\boldsymbol{z})\mathbf{X}$.

The E-M algorithm can be summarized by the following procedure:

Algorithm 1 EM algorithm for the MGP model

Require: $\boldsymbol{\pi}, \mathbf{\Theta}, TOL$
  while $\log p(\mathbf{X}, \mathbf{y}|\mathbf{\Theta}, \boldsymbol{\pi}) - \log p(\mathbf{X}, \mathbf{y}|\mathbf{\Theta}', \boldsymbol{\pi}') < TOL$ **do**
    $\boldsymbol{\pi}' \leftarrow \boldsymbol{\pi}, \mathbf{\Theta}' \leftarrow \mathbf{\Theta}$
    E-Step : Evaluate the expectation $\gamma(\boldsymbol{z})$.
    M-Step : Calculate $\mathbf{\Theta}$ and $\boldsymbol{\pi}$.
  end while

III. Experimental Results

In this section we provide experimental results for the MGP approach. The data consists of agrometeorological variables collected using Wireless Sensor Networks from 5 different locations in the region of Maule in south central Chile[1]. The data was sampled every 10 minutes but only the mean temperature, mean humidity and mean solar radiation values at 15pm and 18pm were used to predict the minimum temperature for next the day. Table I describes the variables used in the predictive model.

x_1	Temp15	Mean Temperature at 15hrs
x_2	Hum15	Mean Humidity at 15hrs
x_3	Rad15	Mean Solar Radiation at 15hrs
x_4	Temp18	Mean Temperature 18hrs
x_5	Hum18	Mean Humidity 18hrs
x_6	Rad18	Mean Solar Radiation at 18hrs
y	$MinNext$	Next day minimum temperature

TABLE I. Variables used for prediction

Each location has several nodes that are used indendently to create the training and testing patterns. In order to demonstrate the ability of the MGP for distributed data processing, in the training data set we incorporate sensor nodes from different locations. Table II describes locations of the sensor nodes used in the training and test datasets. Figure 1 shows the locations in the region of Maule.

Lat	Lon	Site ID
-35.463882	-71.612818	donoso_1
-35.466391	-71.617502	donoso_2
-35.588004	-71.910842	gillmore_2
-35.857414	-71.602135	niceblue_1
-35.858761	-71.601540	niceblue_2
-35.858845	-71.601852	niceblue_3
-35.013470	-71.432918	canepa_1

TABLE II. Locations of the sensor nodes

A number of 100 iterations and a tolerance parameter of 0.0001 were used to train the MGP with $M = 2$ models using Algotithm 1. Figure 2 compares the predicted and observed minimum temperatures at the training sites.

As shown in Figure 2, each cluster produces different predictions. Figure 3 shows the predicted minimum temperatures at the test site. Figure 3 shows the predicted minimum temperatures at a single test site.

[1]http://www.agrosense.cl

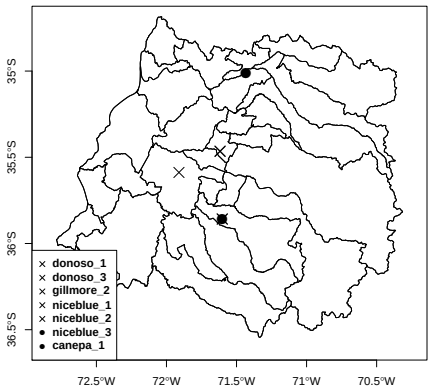

Fig. 1. Locations of the sensor nodes used in the traning and test dataset

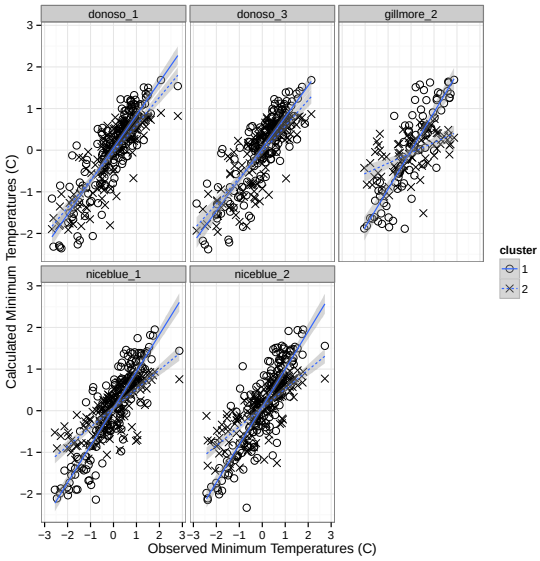

Fig. 2. Comparison between observed and calculated next day minimum temperatures at training locations

Table III summarizes results using the Root Mean Square (RMS) error for the 3 test sites. It is important to notice that the ensemble RMS error is lower for all test locations.

Site ID	RMS Cluster 1	RMS Cluster 2	RMS Ensemble Mean
niceblue_3	0.4314506	0.4521971	0.3712023
canepa_1	0.5649653	0.6597936	0.5737617
canepa_2	0.7197888	0.6106701	0.5600180
canepa_3	0.5906937	0.6441202	0.5677543

TABLE III. RMS values for the MGP model at test locations

IV. Conclusions

In this paper we combine data from multiple and distributed sources and we are able to predict at previously unseen test locations. A mixture of Gaussian process model is introduced as a method to handle the distributed learning problem. The model has the appealing property of being able to handle

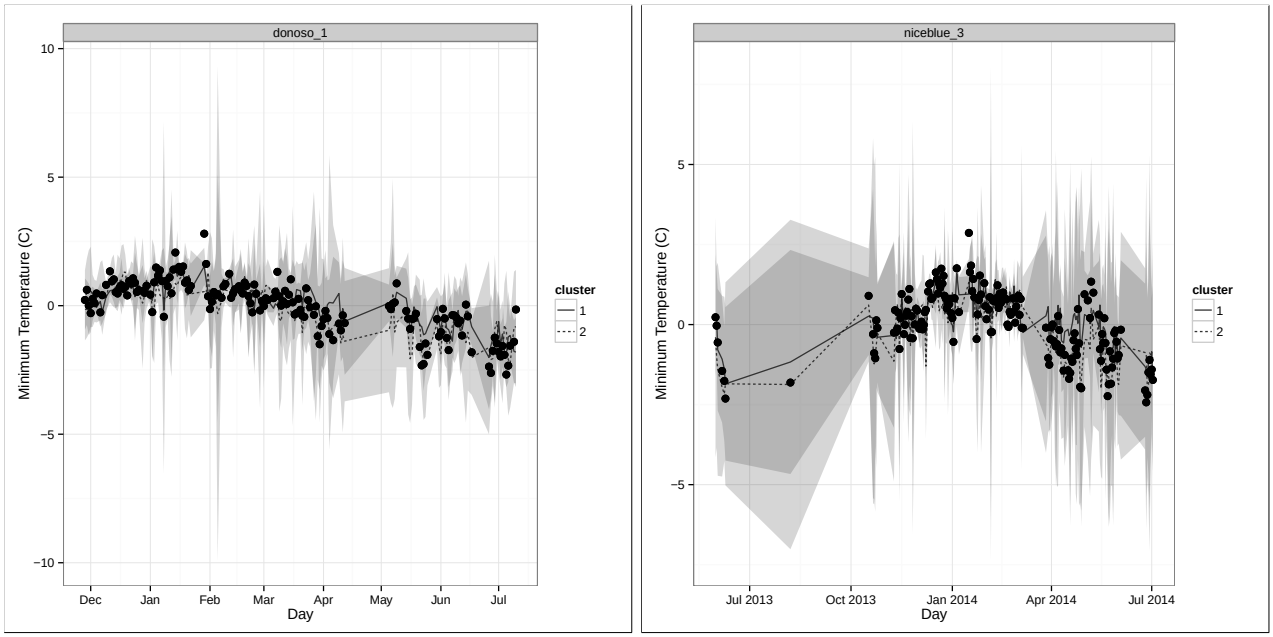

Fig. 3. Minimum temperatures at a training and test locations

multi-modal regression or classification tasks. This is a new approach for distributed learning, since most methods used in the literature use locally trained models and then combine the output. In our case, we overcome the problem of having the same number of models and training locations by imposing the number of mixture components. However, this is also one of the limitations of this approach since we need to choose apriori the number of components. Future work will involve automatic methods for model selection and model averaging procedures for the MGP model.

REFERENCES

[1] R. F. Chevalier, G. Hoogenboom, R. W. McClendon, and J. A. Paz. Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks. *Neural Comput. Appl.*, 20(1):151–159, feb 2011.

[2] R. F. Chevalier, G. Hoogenboom, R. W. McClendon, and J. O. Paz. A web–based fuzzy expert system for frost warnings in horticultural crops. *Environmental Modelling & Software*, 35(0):84 – 91, 2012.

[3] L. Ghielmi and E. Eccel. Descriptive models and artificial neural networks for spring frost prediction in an agricultural mountain area. *Computers and Electronics in Agriculture*, 54(2):101 – 114, 2006.

[4] A Jain, RW McClendon, G Hoogenboom, and R Ramyaa. Prediction of frost for fruit protection using artificial neural networks. *American Society of Agricultural Engineers, St. Joseph, MI, ASAE Paper*, pages 03–3075, 2003.

[5] M. I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.

[6] R. Neal. Regression and classification using Gaussian process priors. In *Bayesian Statistics 6: Proceedings of the sixth Valencia international meeting*, volume 6, page 475, 1998.

[7] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. *Advances in neural information processing systems*, 2:881–888, 2002.

[8] J. Q. Shi, R. Murray-Smith, and D.M. Titterington. Hierarchical gaussian process mixtures for regression. *Statistics and Computing*, 15(1):31–41, 2005.

[9] B. A. Smith, G. Hoogenboom, and R. W. McClendon. Artificial neural networks for automated year-round temperature prediction. *Computers and Electronics in Agriculture*, 68(1):52 – 61, 2009.

[10] R. L. Snyder and J de Melo-Abreu. Frost forecasting and monitoring. *Frost Protection: Fundamentals, Practice, and Economics*, 1:91–112, 2005.

[11] V. Tresp. Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 654–660, 2001.

[12] N. Wang, N. Zhang, and M. Wang. Wireless sensors in agriculture and food industry–recent development and future perspective. *Computers and Electronics in Agriculture*, 50(1):1 – 14, 2006.

[13] Y. Yang and J. Ma. An efficient EM approach to parameter learning of the mixture of gaussian processes. In *Advances in Neural Networks–ISNN 2011*, pages 165–174. Springer, 2011.

[14] C. Yuan, C.and Neubauer. Variational mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 1897–1904, 2009.