Gaussian Process Regression Within an Active Learning Scheme

Conference Paper · July 2011

DOI: 10.1109/IGARSS.2011.6049994 · Source: DBLP

Edoardo Pasolli
Università degli Studi di Trento
35 PUBLICATIONS **333** CITATIONS

Farid Melgani
Università degli Studi di Trento
166 PUBLICATIONS **3,671** CITATIONS

GAUSSIAN PROCESS REGRESSION WITHIN AN ACTIVE LEARNING SCHEME

Edoardo Pasolli and Farid Melgani

Dept. of Information Engineering and Computer Science, Univ. of Trento, Italy
E-mail: pasolli@disi.unitn.it; melgani@disi.unitn.it

ABSTRACT

In this work, we face the problem of training sample collection for the estimation of biophysical parameters by adopting the active learning approach. In particular, we propose two active learning strategies specifically developed for Gaussian Process (GP) regression. The first one is based on adding samples that are distant from the current training samples in the kernel space while the second one exploits an intrinsic GP regression outcome to pick up the most difficult samples. Experiments on simulated and real data sets show the effectiveness of active selection of training samples for regression problems.

Index Terms— Active learning, biophysical parameters, chlorophyll concentration estimation, Gaussian process (GP) regression.

1. INTRODUCTION

One of the most challenging problems in the remote sensing community is represented by the estimation of biophysical parameters from remote sensing data. Several application domains have been considered in the literature such as estimation of biomass concentration in forest areas [1], assessment of ozone concentration in the atmosphere [2], and analysis of water quality through estimation of chlorophyll concentration [3].

From a methodological point of view, several approaches have been proposed to deal with this problem. Because of the strong nonlinearity between biophysical parameters and reflected radiance, nonparametric methods have been preferred to parametric ones despite their greater computational complexity. In particular, strategies based on artificial neural network (ANNs) [4], support vector machines (SVMs) [3], [5] and Gaussian processes (GPs) [6] have been proposed. However, performances of supervised regression approaches depend strongly on the quality and quantity of the data (examples) used to train the regressor. The process of collection of training samples is not trivial, because it is performed manually and thus subject to errors and costs in terms of time and money. For this reason, the number of available training samples is typically limited. A solution to this problem is given by semisupervised approaches, in which the set of training samples is inflated with samples of unknown targets available from the image at zero cost [7]. In more general terms, there is the necessity to find strategies able to choose few training samples, thus minimizing costs, and maintaining high performances in terms of biophysical parameter estimation.

In the field of data classification, a solution to the problem of training sample collection is given by the active learning approach. Starting from a small training set, additional samples are selected from a large amount of unlabeled data. These samples are labeled manually and added to the training set. The process is iterated until a stopping criterion is reached. Active learning strategies have been applied successfully in different remote sensing application fields, such as detection of buried objects [8], classification of hyperspectral images [9], and classification of very high spatial resolution images [10], [11] as well as in the biomedical field such as the classification of ECG beats [12]. Despite the promising performance given by the active learning approach in the classification field, nothing similar has been proposed in the remote sensing community for regression problems.

The objective of this work is to propose some strategies of active learning for regression problems. In particular, solutions for problems based on GP regression are presented. The proposed strategies and the related experiments are discussed in the next Sections.

2. PROPOSED METHOD

Before detailing the proposed active learning strategies, we introduce briefly the key equations for GP regression [13], [14].

2.1. GP regression

In GP theory, the learning of a regressor is formulated in terms of a Bayesian estimation problem, where the predictive function is inferred from the realizations of a random process drawn from a Gaussian distribution. Given a generic unknown sample $\mathbf{x}_*$, the best estimation of the output value f_* associated with it is represented by the expectation of the

978-1-4577-1005-6/11/$26.00 ©2011 IEEE

desired output quantity conditioned to L and $\mathbf{x}_*$:

$$\hat{f}_*|L, \mathbf{x}_* \sim E\left\{f_*|L, \mathbf{x}_*\right\} = \int f_* p\left(f_*|L, \mathbf{x}_*\right) df_*. \tag{1}$$

It can be shown that the predictive distribution $p\left(f_*|L, \mathbf{x}_*\right)$ takes the following expression:

$$p\left(f_*|L, \mathbf{x}_*\right) \sim N\left(\mu_*, \sigma_*^2\right) \tag{2}$$

where

$$\mu_* = \mathbf{k}_*^t \left[K + \sigma_n^2 I\right]^{-1} \cdot \mathbf{y} + b \tag{3}$$

$$\sigma_*^2 = k\left(\mathbf{x}_*, \mathbf{x}_*\right) - \mathbf{k}_*^t \left[K + \sigma_n^2 I\right]^{-1} \mathbf{k}_* \tag{4}$$

and, given a covariance function $k\left(\mathbf{x}, \mathbf{x}'\right)$, K and $\mathbf{k}_*$ represent the covariance matrix of the training samples and the covariance vector between the training samples and the sample $\mathbf{x}_*$, respectively. Moreover b, σ_n and I are the bias factor, the noise variance, and the identity matrix, respectively.

Two important elements can be retrieved from them: 1) the mean μ_* which represents the best output value estimation for the considered sample according to (1); 2) the variance σ_*^2 which expresses a confidence measure associated by the model to the output.

A central role is played by the covariance function $k(\mathbf{x}, \mathbf{x}')$ as it embeds the geometrical structure of the training samples. A typical choice for the covariance function is the squared exponential function:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\left|\mathbf{x} - \mathbf{x}'\right|^2}{2l^2}\right) \tag{5}$$

where the two hyperparameters σ_f^2 and l are called process variance and length scale, respectively.

2.2. Active learning strategies

We consider a training set composed initially of n labeled samples $L = \left\{\mathbf{x}_i, y_i\right\}_{i=1}^n$, where $\mathbf{x}_i$ represents a vector of remote observations and y_i is the associated target value. Additionally, we consider a learning set of m samples $U = \left\{\mathbf{x}_j\right\}_{j=n+1}^{n+m}$, with $m >> n$, for which the corresponding target values are unknown. In the following, two different strategies of active learning for GP regression are proposed.

The first strategy (TRkd) is based on the distances in the feature space from the samples already composing the current training set. In particular, it consists to calculate for each sample $\mathbf{x}_j$ $(j = n+1, n+2, \cdots, n+m)$ the Euclidean distances from the training samples $\mathbf{d}_j \in \Re^{TrS} = \left[d_{j,1}, d_{j,2}, \cdots, d_{j,TrS}\right]$:

$$d_{j,i} = \left\|\mathbf{x}_j - \mathbf{x}_i\right\| \tag{6}$$

where *TrS* is the current number of training samples. After calculating the distances from the training samples, these distance values are combined opportunely through the same kernel operator used by the GP regressor. The criterion of selection $f_{TRkd,j}$ for the sample $\mathbf{x}_j$ is given by the following formulation:

$$f_{TRkd,j} = \sum_{i=1}^{TrS} k\left(d_{j,i}\right) \tag{7}$$

where $k\left(d_{j,i}\right)$ is the squared exponential function defined in (5). The function f_{TRkd} assumes lower values when the considered samples are further from the training samples. Therefore, the samples characterized by the lower values of f_{TRkd} are selected. In this way, we favor the selection of sample placed in areas of the feature space not covered by training samples and avoid to choose samples similar to those already present in the current training set.

The second strategy (VoP) is based on the measure of variance on the predictions defined in (4). This value expresses a confidence measure associated by the model to the output and therefore provides an information on the reliability of the estimations. The criterion of selection $f_{VoP,j}$ for the sample $\mathbf{x}_j$ is given by the following formulation:

$$f_{VoP,j} = \sigma_*^2\left(\mathbf{x}_j\right) \tag{8}$$

where $\sigma_*^2\left(\mathbf{x}_j\right)$ is the variance measure defined in (4). The function f_{VoP} tends to zero when the confidences on the estimations are high. Since we desire to enrich the current training set with new and difficult samples, we choose the samples with the greater values of f_{VoP}.

3. EXPERIMENTS

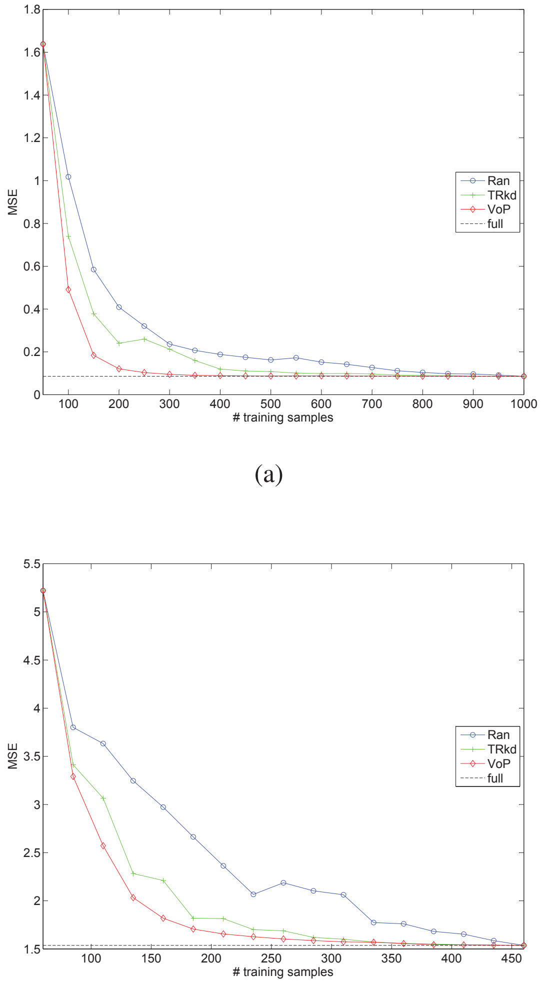
3.1. Experimental setup

In order to validate the proposed strategies we have conducted an experimental study on simulated and real data sets.

The first data set refers to multispectral data that simulate the spectral behavior of the chlorophyll concentration in subsurface case I + case II (open and costal) waters, through the first eight channels (412-618 nm) of the multispectral Medium Resolution Imaging Spectrometer (MERIS) satellite sensor. The range of variation of the chlorophyll concentration is from 0.02 to 54 mg/m3.

The second data set is the SeaWiFS Bio-optical Algorithm Mini-Workshop (SeaBAM) one. It represents real measurements of chlorophyll concentration, mostly in case I (open) waters, around the U.S. and Europe related to five different Sea-viewing Wide Field-of-view Sensor (SeaWiFS) wavelengths (412, 443, 490, 510, and 555 nm). The chlorophyll concentration values span an interval between 0.02 and 32.79 mg/m3.

All the available samples were split in two sets, corresponding to learning set U and test set. The initial training samples were selected randomly from the learning set U. For the MERIS data set, starting from 50 samples, the active learning algorithms were run until all the learning samples were added to the training set, adding 50 samples at each iteration. Similarly, for the SeaBAM data set, 25 samples were



(a)

(b)

Fig. 1. Performances in terms of MSE achieved on (a) the MERIS and (b) the SeaBAM data set, respectively. Each graph shows the results in function of the number of training samples and averaged over ten runs of the algorithm, each with a different initial set. Ran= random, TRkd= distances from training samples combined with kernel operator, VoP=variance on predictions, full= full GP.

added at each iteration by starting from 60 samples. The entire active learning process was run ten times, each time with a different initial training set to yield statistically reliable results. At each run, the initial training samples were chosen in a completely random way.

A GP regressor was also trained on the entire learning set in order to have a reference-training scenario, called ”full” training. On one hand, the regression results obtained in this way represent an upper bound for the accuracies. On the other hand, we expect that the lower accuracy bound will be given by the completely random selection strategy (Ran). We recall that the purpose of any active learning strategy is to converge to the performance of the ”full” training scenario faster than the Ran method. Regression performance was evaluated in terms of mean squared error (MSE).

3.2. Experimental results

The regression performances are shown in Fig. 1(a),(b). In particular, the values of MSE as a function of the number of selected training samples are depicted for the MERIS and the SeaBAM data set, respectively. We note that the active selection of the training samples allows a faster convergence to the ”full” accuracy with respect to the random strategy. In particular, the strategy VoP shows the best performances. For the MERIS data set, this strategy converges to the ”full” accuracy using about 300 training samples, which represent 30% of the entire learning set. Instead, the entire set of training samples is necessary for the Ran method to converge. Similarly, for the SeaBAM data set, about 310 samples are required to converge for the proposed strategies. Additionally, the proposed methods give best performances before convergence for both data sets. This means that similar values of accuracies can be obtained using a minor quantity of training samples, which implies a reduction of the manual work for giving target values to samples and a decreasing of the computational time necessary to train the regressor.

The obtained results in terms of MSE are shown in greater detail in Table 1(a),(b). In particular, the proposed strategies are evaluated at convergence, i.e. 300 and 310 samples are considered for the MERIS and the SeaBAM data set, respectively. It is confirmed how the proposed strategies are characterized by better performances with respect to the random selection.

4. CONCLUSION

In this work, we have investigated the active learning approach in the process of training sample collection for regression problems related to the estimation of biophysical parameters from remote sensing data. In particular, we have proposed two strategies specifically developed for GP regression. They are based on distances from training samples combined with kernel operator and variance measure on the predictions. The obtained experimental results show promising capabilities of the proposed strategies in terms of training sample collection. While in this work we focused on GP regression, research is in progress in order to use active learning approach in combination with other supervised regression approaches.

5. REFERENCES

[1] D.G. Goodenough, AS Bhogall, H. Chen, and A. Dyk, “Comparison of methods for estimation of Kyoto pro-

TABLE I. MSE Achieved on (a) the MERIS and (b) the SeaBAM Data Sets

(a)

Method	# training samples	MSE
Full	1000	0.086
Initial	50	1.638
R	300	0.237
TRkd		0.212
VoP		0.095

(b)

Method	# training samples	MSE
Full	460	1.536
Initial	60	5.221
R	310	2.062
TRkd		1.601
VoP		1.573

tocol products of forests from multitemporal Landsat," in *Proc. IGARSS*, Sidney, AUS, Jul. 2001, IEEE, vol. 2, pp. 764–767.

[2] F. Del Frate, A. Ortenzi, S. Casadio, and C. Zehner, "Application of neural algorithms for a real-time estimation of ozone profiles from GOME measurements," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2263–2270, Oct. 2002.

[3] L. Bruzzone and F. Melgani, "Robust multiple estimator systems for the analysis of biophysical parameters from remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 1, pp. 159–174, Jan. 2005.

[4] D. D'Alimonte and G. Zibordi, "Phytoplankton determination in an optically complex coastal region using a multilayer perceptron neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2861–2868, Dec. 2003.

[5] G. Camps-Valls, L. Bruzzone, J.L. Rojo-Álvarez, and F. Melgani, "Robust support vector regression for biophysical variable estimation from remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 3, pp. 339–343, Jul. 2006.

[6] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 3, pp. 464–468, Jul. 2010.

[7] Y. Bazi and F. Melgani, "Semisupervised PSO-SVM regression for biophysical parameter estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1887–1895, Jun. 2007.

[8] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: Application to sensing subsurface UXO," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 11, pp. 2535–2543, Nov. 2004.

[9] S. Rajan, J. Ghosh, and M.M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[10] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[11] E. Pasolli, F. Melgani, and Y. Bazi, "Support vector machine active learning through significance space construction," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 431–435, May 2011.

[12] E. Pasolli and F. Melgani, "Active learning methods for electrocardiographic signal classification," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 6, pp. 1405–1416, Nov. 2010.

[13] C.K.I. Williams and C.E. Rasmussen, "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems*. 1996, vol. 8, pp. 514–520, Cambridge, MA: MIT Press.

[14] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press, 2006.