SEMISUPERVISED GAUSSIAN PROCESS REGRESSION FOR BIOPHYSICAL PARAMETER ESTIMATION

Yakoub Bazi and Farid Melgani***

* College of Engineering, King Saud University, Riyadh, Saudi Arabia
E-mail: yakoub.bazi@ju.edu.sa

** Dept. of Information Engineering and Computer Science, Univ. of Trento,
Via Sommarive, 14, I-38050 Trento, Italy
E-mail:melgani@disi.unitn.it

PRESENTATION: SPECIAL SESSION ON "MACHINE LEARNING AND REMOTE SENSING"

ABSTRACT

In the last years, the growing importance of a large scale understanding and monitoring of the earth system and the rapid sensor technology development have animated a strong interest of the remote sensing community to the problem of estimating biophysical parameters from remote sensing data. Examples of related applications are the estimation of ozone concentration in the atmosphere, biomass concentration in forest areas, and water quality parameters like the chlorophyll concentration for monitoring oceans and costal areas.

From a methodological point of view, this problem can be approached by viewing it as an inverse modeling issue which can be solved in two different ways. These last consist to define a model which relates the acquired observations to the parameter of interest. While the first way is based on a parametric model which depends on a predefined set of parameters (e.g., polynomial and exponential models), the second one makes use of a nonparametric model whose behavior is completely data-dependent. In both cases, the model is estimated by regression from training samples (pairs of received radiances and in-situ measurements of the biophysical parameter concentration). Usually, parametric models are not sufficiently effective to capture the relationships between biophysical parameters and acquired radiances due to their complexity. By contrast, nonparametric techniques often prove to be more suitable for performing such task, but at the price of higher computational complexity. Among these techniques, one can find artificial neural networks (ANNs) [1] and support vector machines (SVMs) [2], [3].

Recently, a new machine learning approach that is based on the Gaussian process (GP) theory has been introduced in the literature and has shown particularly promising for estimating biophysical parameters from remotely sensed data [4]-[5]. According to this approach, the learning of a machine (regressor or classifier) is formulated in terms of a Bayesian estimation problem, where the parameters of the machine are assumed to be

random variables which are a-priori jointly drawn from a Gaussian distribution. The underlying idea of the GP regression can be described in different ways. The simplest one consists to adopt a linear regression model in which the parameters are assumed to be jointly Gaussian random variables with a predefined mean vector (generally the origin of the parameter space) and covariance matrix. The observed values of the linear function (to model) are supposed to come from the sum of the (latent) linear function and a Gaussian noise. The estimation of the linear model parameters is carried out under a Bayesian framework which requires the estimation of the posterior probability function over the parameters. It is shown that the posterior function is a Gaussian distribution whose parameters depend on the training samples, the covariance matrix and the noise variance. As a result, the probability distribution of the function observations follows a Gaussian behavior whose mean and variance depend also on the input feature vector (for which the function estimate is desired). Since most of the real regression problems are not expected to be linear, the above regression procedure is reformulated by simply kernelizing the linear regression model. Consequently, the distribution of the function observations becomes a function of kernel distances between training samples, defining the kernel matrix, and between the training samples and the sample for which the estimation is desired. Two important information can be retrieved from this distribution: 1) the mean, which will represent the function (output) estimate for the considered sample; and 2) the variance which will quantify the confidence associated whit the output estimate.

In general, however, the success of the above mentioned regression methods is conditioned by the availability of sufficient and representative training samples to obtain reliable estimation accuracies. Nonetheless, in practice collecting a statistically significant and representative amount of training (labeled) samples is often a difficult and time consuming task. To face such issue, a semisupervised regression method based on SVM regression has been introduced recently [6]. Its underlying idea is to inflate the training set with unlabeled samples, which are readily available at zero cost from the remote sensing data under analysis, to compensate the scarcity of labeled samples. By unlabeled data, we mean generic samples whose input (observation) values are known, whereas the corresponding desired outputs (biophysical parameter concentrations) are unknown.

In this work, we propose to apply a similar inflation approach to the GP regression problem. In particular, during the learning phase of the approach, unlabeled samples are exploited to inflate the training set. The estimation of the targets associated with these samples is carried out by solving an optimization problem formulated within a genetic optimization framework [7]. The search process of the target estimates is guided by the separate or joint optimization of two different criteria expressing the generalization capabilities of the GP estimator. The two explored criteria are: 1) the empirical risk quantified in terms of the mean square error (MSE) measure; and 2) the log marginal likelihood. This last merges two terms expressing the model complexity and the data fit capability, respectively. The joint optimization is performed by means of the multiobjective nondominated sorting genetic algorithm (NSGA-II) [8]. The selection of the unlabeled samples has been envisioned according to

three different criteria: random sampling, estimate variance provided by the GP regressor and differential entropy measure [9].

Several experiments were conducted on simulated as well as real datasets referring to measurements of chlorophyll concentration in coastal waters. The obtained results show that significant gains of estimation accuracy can be achieved thanks to the inflation process in particular when adopting a multiobjective optimization scheme with a selection of the unlabeled samples based on the differential entropy measure.

Keywords: Biophysical parameter estimation, Gaussian processes, genetic algorithms, multiobjective optimization, regression methods, semisupervised learning.

References

[1] P. Cipollini, G: Corsini, M. Diani, R. Grosso, "Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks.", *IEEE Trans. Geosci. and Remote Sens.*, vol. 39, pp. 1508-1524, 2001.

[2] L. Bruzzone, F. Melgani, "Robust multiple estimator systems for the analysis of biophysical parameters from remotely sensed data." *IEEE Trans. Geosci. and Remote Sens.*, vol. 43, pp.159-174, 2005.

[3] H. Zahn, P. Shi, C. Chen, "Retrieval of oceanic chlorophyll concentration using support vector machines.", *IEEE Trans. Geosci. and Remote Sens.*, vol. 41, pp. 2947-2951, 2003.

[4] C.E. Rasmussen and C.K.I. Williams, *Gaussian Process for machine learning*. Cambridge, Massachusetts: The MIT Press, 2006.

[5] L. Pasolli, F. Melgani, and E. Blanzieri, "Estimating Biophysical Parameters from Remotely Sensed Imagery with Gaussian Processes", Proc. of the *IEEE-International Geoscience and Remote Sensing Symposium IGARSS-2008*, Boston, USA, vol. 2, pp. 851-854, July 2008.

[6] Y. Bazi and F. Melgani, "Semisupervised PSO-SVM regression for biophysical parameter estimation." *IEEE Trans. Geosci. and Remote Sens.*, vol. 45, pp. 1887-1895, 2007.

[7] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[9] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.