Prediction of Data Stream Based on Gaussian Process Regression with Online Free Variational Inference Approximation

IOP Conf. Series: Materials Science and Engineering **569** (2019) 052056 doi:10.1088/1757-899X/569/5/052056

Prediction of Data Stream Based on Gaussian Process Regression with Online Free Variational Inference Approximation

Xin Wang[1,2*], Jiangyong Duan[1] and Zhen Yan[1]

[1] Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, 100094, China

[2] University of Chinese Academy of Sciences, Beijing, 100049, China

[*]Corresponding author's e-mail: wangxin16@csu.ac.cn

Abstract. Online prediction of data stream is currently being used in various fields. The prediction method based on Gaussian Process Regression has obvious advantages since its outputs have a probabilistic significance, which is suitable for dealing with nonlinear and complex regression problem. However, the characteristics of the data stream are real-time and online. If only using the historical data of the previous time as the training sample to construct the prediction model, the model will not be accurately predicted once the distribution of the new data changes. To this end, we employ the Online Free Variational Inference approximation to build the prediction model. The key idea is to introduce a variational distribution and maximizing the Kullback-Leibler Divergence between the variational distribution and the true value of the output's posterior distribution. Further, we use telemetry data to verify the validity of the Online Free Variational Inference approximation and the experiment shows that the online data stream can be predicted well by the method employed in the paper.

1. Introduction

Space science experiments and space observation activities have gradually increased nowadays. The payload application system is one of the most important sub-systems in the space missions, especially space science experiments. In order to ensure the normal operation of each payload, how to find faults in time and solve the faults quickly and unambiguously are one of the problems that ground-based receivers and researchers have been studying.

The basis for judging the working state of each payload is the downlink telemetry data of each payload. If the data can be analyzed and processed effectively and accurately, and we can make good predictions of the data arriving in the future, it will be of great help to the follow-up experiments. Gaussian Process Regression can be better used for prediction because it can get the probability of output [1]. Therefore, this method is used more and more widely.

The non-parametric of Gaussian Process Regression leads to its huge computational complexity. In order to solve this problem, researchers have studied some sparse methods, such as Subset Data approximation, Sparse Pseudo-inputs GP, etc. in references [2~9]. Most of these methods are based on historical data for a period of time to build a prediction model. However, if we continue to use the previous historical data as training datasets, the accuracy of the prediction may be greatly reduced once the data distribution at the next time period has changed. Therefore, an online Gaussian Process

Regression algorithm is needed to construct an online prediction model in order to ensure that the data of each time period can be predicted in real time.

In this paper, we use the Online Free Variational Inference approximation (OFVI) to construct a model to realize the prediction of the online data stream. The main idea is to introduce a variational distribution to ensure that the variational distribution function and the true value of the output's posterior distribution are as close as possible by maximizing the Kullback-Leibler Divergence (KL divergence). In order to reduce the time complexity, we introduce the inducing points to perform sparse approximation on the input data. Further, we use the gradient descent method to optimize the hyperparameters to obtain the optimal model for the prediction of data stream.

2. Online Gaussian Process Regression with the free variational inference approximation

2.1. Background of the free variational inference approximation

Gaussian Process is a set of arbitrary finite random variables with a joint Gaussian distribution which properties are determined by the mean function and the covariance function [10]. Given N input and output pairs $\{x_n, y_n\}_{n=1}^N$, assuming noise, and defining Gaussian noise is $\varepsilon_n \sim N(0, \sigma_y^2)$, the relationship between input and output is $y_n = f(x_n) + \varepsilon_n$ where $f \sim GP(m(x), K(\cdot,\cdot \mid \theta))$. We can calculate the posterior over f, $p(f \mid y, \theta)$ and the marginal likelihood $p(y \mid \theta)$ from these formulas [11]. However, it involves non-linear and unconstrained extreme problem, and involves inversion of the covariance matrix of $n \times n$ dimensions every gradient calculation in the process of optimizing hyperparameters. Therefore, the calculation amount of the training process of Gaussian Process Regression (GPR) is $O(n^3)$, and the calculation amount of the covariance prediction is $O(n^2)$.

The computational efficiency is very low when the data set is huge. In order to solve this problem, it is necessary to save Gaussian Process (GP) training time through sparse approximation and so on, thereby reducing time complexity and computational complexity. In this paper, we use the free variational inference approximation to reduce the time complexity.

Given a variational distribution $q(f)$, we can get the logarithm of $p(y \mid \theta)$ by equation (1):

$$\log p(y \mid \theta) = \log \int \mathrm{d}\, f p(y, f \mid \theta) \geq \int \mathrm{d}\, f q(f) \log \frac{p(y, f \mid \theta)}{q(f)} = F_{fvi}(q, \theta) \tag{1}$$

where

$$F_{fvi}(q, \theta) = \log p(y \mid \theta) - KL[q(f) \parallel p(f \mid y, \theta)] \tag{2}$$

From equation (2), we can see that it is possible to ensure $q(f)$ gets closer to the exact posterior $p(f \mid y, \theta)$ as close as possible by maximizing the KL divergence. In order to reduce the amount of calculation, we select a set of u data set as the inducing points, so the approximate posterior distribution can be set by equation (3):

$$q(f) = p(f_{\neq u} \mid u, \theta) q(u) \tag{3}$$

where $q(u)$ is a variational distribution over u and $p(f_{\neq u} \mid u, \theta)$ is the prior distribution of the remaining latent function values. Therefore we can get $F_{fvi}(q(u), \theta)$ by equation (4):

$$\begin{aligned} F_{fvi}(q(u), \theta) &= \int \mathrm{d}\, f q(f) \log \frac{p(y \mid f, \theta) p(u \mid \theta) p(f_{\neq u} \mid u, \theta)}{p(f_{\neq u} \mid u, \theta) q(u)} \\ &= -KL[q(u) \parallel p(u \mid \theta)] + \sum_n \int \mathrm{d}\, u q(u) p(f_{x_n} \mid u, \theta) \log p(y_{x_n} \mid f_{x_n}, \theta) \end{aligned} \tag{4}$$

According to equation (1) ~ equation (4), we can deduce $q_{fvi}(f)$ and $F_{fvi}(\theta)$ by equation (5) and equation (6):

$$p(f \mid y, \theta) \approx q_{fvi}(f) \propto p(f_{\neq u} \mid u, \theta) p(u \mid \theta) N(Y; K_{fu} K_{uu}^{-1} u, \sigma_y^2 I) \tag{5}$$

$$\log p(y \mid \theta) \approx F_{fvi}(\theta) = \log N(y; 0, K_{fu} K_{uu}^{-1} K_{uf} + \sigma^2 I) - (2\sigma_y^2)^{-1} \sum_n (k_{nn} - K_{nu} K_{uu}^{-1} K_{un}) \tag{6}$$

After the above approximation, the time complexity for approximate maximum likelihood learning becomes $O(nu^2)$. The above derivation process introduces the method of free variational inference approximation. The goal of this paper is to achieve prediction of online data stream, so we will introduce the idea of Online Free Variational Inference approximation in the next subsection, and using this idea to construct a model and optimize it to achieve prediction of data stream.

2.2. Online free variational inference approximation

The object of this paper is the telemetry data stream returned by the payload, and the data arrives in order. The goal is to ensure that data is predicted in time when the new time period arrives. The amount of calculation is huge if using all the historical data to build a prediction model, which leads to low efficiency [12]. Therefore, we access the data points of the current time period only. Further, the effect of the old data on the current posterior distribution needs to be propagated through the posterior distribution of the previous time period. Here, we introduce the inducing points that can represent the old data to reduce the calculation complexity. The inducing points also need to be adjusted online since the new parts will come over time. We will introduce the Online Free Variational Inference approximation (OFVI) and the optimization process of hyperparameters as follow.

Define the true posterior distribution of the previous time period is $q_{old}(f)$, and the posterior distribution of the new arrival time period is $q_{new}(f)$, where f representing the function of training data. We can get equation (7) and equation (8):

$$q_{old}(f) \approx p(f \mid y_{old}) = \frac{1}{z_1(\theta_{old})} p(f \mid \theta_{od}) p(y_{old} \mid f) \tag{7}$$

$$q_{new}(f) \approx p(f \mid y_{old}, y_{new}) = \frac{1}{z_2(\theta_{new})} p(f \mid \theta_{new}) p(y_{old} \mid f) p(y_{new} \mid f) \tag{8}$$

and we can get an approximation of $p(y_{old} \mid f)$ by equation (7):

$$p(y_{old} \mid f) \approx \frac{z_1(\theta_{old}) q_{old}(f)}{p(f \mid \theta_{old})} \tag{9}$$

Substituting equation (9) into equation (8), an approximation of $p(f \mid y_{old}, y_{new})$ can be obtained as equation (10):

$$\hat{p}(f \mid y_{old}, y_{new}) = \frac{z_1(\theta_{old})}{z_2(\theta_{new})} p(f \mid \theta_{new}) p(y_{new} \mid f) \frac{q_{old}(f)}{p(f \mid \theta_{old})} \tag{10}$$

From equation (10), if the hyperparameter is a fixed value, the posterior distribution of the new data point can be derived. However, it is more difficult to find the result when the hyperparameter is updated online. In this paper, we use the inducing points to update the distribution of the variation. Here, we allow the location of the inducing input points in the approximation of the new data to be different from those in the old data.

Define $a = f(z_{old})$ represents the inducing input points of the previous time period, and $Num_a = |a|$ represents the number of the inducing input points. At the same time, $b = f(z_{new})$ is defined to represent the new incoming inducing input points for the next time period, and $Num_b = |b|$ represents the number of its inducing points. Assuming $q_{old}(a)$ obeys Gaussian distribution, that is, $q_{old}(a) = N(a; M_a, K_a)$, so the true posterior distribution $q_{old}(f)$ of the previous time period can be expressed as equation (11):

$$q_{new}(f) = p(f_{\neq b} \mid b, \theta_{new}) q_{new}(b) \tag{11}$$

Similarly, the posterior distribution of the new coming data can be expressed as equation (12). Further, its inducing points and hyperparameters are all updated.

$$q_{new}(f) = p(f_{\neq b} \mid b, \theta_{new}) q_{new}(b) \tag{12}$$

Therefore, the approximate inference problem is transformed into the optimization problem using variational inference after processing. According to the principle of the variational inference

approximation algorithm introduced in the previous subsection, the KL divergence of $q_{new}(f)$ and $\hat{p}(f \mid y_{old}, y_{new})$ can be expressed as equation (13):

$$KL[q_{new}(f) \| \hat{p}(f \mid y_{old}, y_{new})] = \int \mathrm{d}\, f q_{new}(f) \log \frac{p(f_{\neq b} \mid b, \theta_{new}) q_{new}(b)}{\frac{z_1(\theta_{old})}{z_2(\theta_{new})} p(f \mid \theta_{new}) p(y_{new} \mid f) \frac{q_{old}(f)}{p(f \mid \theta_{old})}}$$
$$= \log \frac{z_2(\theta_{new})}{z_1(\theta_{old})} + \int \mathrm{d}\, f q_{new}(f) \left[\log \frac{p(a \mid \theta_{old}) q_{new}(b)}{p(b \mid \theta_{new}) q_{old}(a) p(y_{new} \mid f)} \right] \quad (13)$$

The next step is to optimize the hyperparameters to get the optimal solution.

In order to optimize the hyperparameters, we calculate the boundary value of the online negative logarithmic marginal likelihood for equation (13). Deduce $F = (q_{new}(f), \theta_{new})$ and let $q(b)$ be equal to 0, we can get the optimal approximate posterior expressed as equation (14):

$$q_{ofvi}(b) \propto p(b) \exp\left(\int \mathrm{d}\, a p(a \mid b) \log \frac{q_{old}(a)}{p(a \mid \theta_{old})} + \int \mathrm{d}\, f p(f \mid b) \log p(y_{new} \mid f) \right) \quad (14)$$
$$\propto p(b) N\left(\hat{y}; K_{\hat{fb}} K_{bb}^{-1} b, \sum\nolimits_{\hat{y}, ofvi}\right)$$

where f is the function of the new training points and

$$\hat{y} = \begin{bmatrix} y_{new} \\ D_a S_a^{-1} m_a \end{bmatrix}, \quad K_{\hat{fb}} = \begin{bmatrix} K_{fb} \\ K_{ab} \end{bmatrix}, \quad \sum\nolimits_{\hat{y}, ofvi} = \begin{bmatrix} \sigma_y^2 I & 0 \\ 0 & D_a \end{bmatrix}, \quad D_a = \left(S_a^{-1} - K_{aa}'^{-1}\right)^{-1}$$

Deriving θ, we can get equation (15):

$$F(\theta) = \log N\left(\hat{y}; 0, K_{\hat{fb}} K_{bb}^{-1} K_{b\hat{f}} + \sum\nolimits_{\hat{y}, fvi}\right) - \frac{1}{2\sigma_y^2} tr\left(K_{ff} - K_{fb} K_{bb}^{-1} K_{bf}\right) + \Delta_a \quad (15)$$

where

$$\Delta_a = \left(-\log|S_a| + \log|K_{aa}'| + \log|D_a| + m_a^T \left(S_a^{-1} D_a S_a^{-1} - S_a^{-1}\right) m_a - tr\left[D_a^{-1} Q_a\right] + const\right) / 2$$

Equations (14) and (15) represent the process of hyperparameter adaptive learning in the Online Free Variational Inference approximation. The algorithm will be validated using the telemetry data stream of the on-track payload in Section 3.

3. Experiment

In this section, the downlink telemetry data stream returned by the temperature sensor of a certain load on the rail is used as the data set to test the feasibility of the Online Free Variational Inference approximation (OFVI) described.

The data in the experiment is arrived in chronological order and we select the two columns with high correlation. The mean function is set zero. The Gaussian Covariance function is the squared exponential with isotropic distance measure. The optimal hyperparameters are obtained by minimizes the negative log marginal likelihood in gradient ascent method. For the convenience of comparison, we put the data coming from different time period into one picture, as shown in Figure 1.

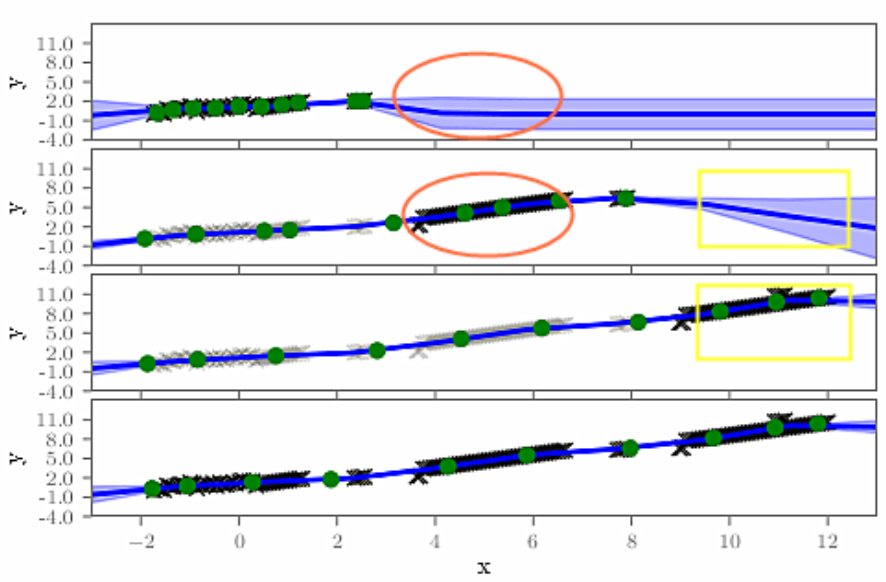

Figure 1. Online prediction using the online free variational inference approximation

In Figure 1, the black dots represent the new arrival data, the gray dots represent the old data of the previous time period, the blue curve represents the calculated prediction mean, and the blue shaded areas are the predicted confidence intervals. The green dots represent the set of inducing input points, which are distributed according to the characteristics of the data. The first row in Figure 1 shows the mean and the confidence interval based on the prediction model by using the data of the current time period. In the second row, the gray dots on the left represent the data of the previous time period, and the black dots on the right represent the new arrival data. Comparing the area of the red circle in the first row and the second row of Figure 1, the confidence interval originally predicted in the first row is relatively wider, but the distribution of the real data points obtained in the second row is inconsistent with the first one. Therefore, it indicates that the distribution of the newly arrived data does not follow the distribution of the old data in the previous time period, so it is not accurate to construct the prediction model only using the historical data as the training sample, and it is necessary to consider the prediction model for online updating.

Similarly, in the third row, the gray data points of the two areas on the left represent the data that have flowed in the first two old periods, and the black points on the right represent the latest time period of the new arrival data. The area enclosed by the yellow rectangle in the second row represents the mean and confidence interval predicted based on the historical data coming from the previous one. The area enclosed by the yellow rectangle in the third row represents the distribution of the real data coming from the current time period. Comparing these two rows, we can found that the distribution of new incoming data does not follow the previous historical data distribution, so the online method ensures the accuracy of the prediction.

The fourth row is a prediction model using all the datasets that come from the first three rows. It can be seen that the distribution shown is basically the same as it in the third one, thus proving the correctness and feasibility of the method described in the paper. However, since the fourth row constructs the prediction model with all the data as training samples, the complexity will increase greatly, so the effect in practical applications is not very satisfactory. In our experiment, Mean Square Error (MSE), R-square and time are used as model performance evaluation criteria. The smaller the Mean Squared Error, and the larger the R-square, the better the model's generalization ability and the higher the model's prediction accuracy. Table 1 shows the performance comparison results of the construction of the prediction model using the Full GP and the Online Free Variational Inference approximation (OFVI).

Table 1. Comparision of prediction model between using Full GP and Online Free Variational Inference approximation (OFVI).

Method	Time/s	MSE	R-square
Full GP	7823.681	0.00462	0.89948
OFVI	12.265	0.00471	0.89932

Here, we can see that the training time using the Online Free Variational Inference approximation (OFVI) is significantly reduced by about 600 times compared with Full GP, but MSE and R-square have almost not changed. Therefore, the method we used is reasonable and available.

4. Conclusion

We have introduced the learning framework of the Online Gaussian Process Regression. Since the arrival of the data stream is continuous, if only using the historical data of a period time as the training sample for prediction, the accuracy of the prediction will be greatly reduced once the subsequent arrival data does not satisfy the distribution of the data at the previous time period. Therefore, this paper adopted the Online Variational Free Inference approximation, which can realize real-time update learning, and the predicted value of each period is obtained from the training samples of the latest. In addition, the computational complexity is greatly reduced and the operation efficiency is improved by introducing the inducing points. We use the downlink telemetry data stream of the on-rail payload to carry out experiments, and the feasibility and effectiveness of the method are verified. The method in this paper is suitable for real-time data stream and is increasingly used in the life. In the future, we will further optimize the algorithm and apply it to more fields of data stream.

References

[1] Cao, Y. , A. Brubaker, M. , J. Fleet, D. , & Hertzmann, A. . (2015). Efficient optimization for sparse gaussian process regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(12), 2415-2427.

[2] Sjolund, J. , Eklund, A. , Ozarslan, E. , & Knutsson, H. . (2017). Gaussian process regression can turn non-uniform and undersampled diffusion MRI data into diffusion spectrum imaging. IEEE International Symposium on Biomedical Imaging. IEEE. pp. 778-782.

[3] Liang, K. , Chahir, Y. , Molina, Michèle, Tijus, C. , & Jouen, F. . (2013). [acm press the 2013 conference - cape town, south africa (2013.08.29-2013.08.31)] proceedings of the 2013 conference on eye tracking south africa - etsa \"13 - appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression. 17-23.

[4] Bauer, M. , Mark, V. D. W. , & Rasmussen, C. E. . (2016). Understanding probabilistic sparse gaussian process approximations. pp. 1525-1533.

[5] Park, J. , Law, K. H. , Bhinge, R. , Chen, M. , & Rachuri, S. . (2015). Real-Time Energy Prediction for a Milling Machine Tool Using Sparse Gaussian Process Regression. In: IEEE International Conference on BigData. CA, USA. pp. 1201-1209.

[6] Hombal, V. , & Mahadevan, S. . (2011). Bias minimization in gaussian process surrogate modeling for uncertainty quantification. Int.j.uncertain.quantif, 1(1), 321-349.

[7] Nguyentuong, D. , Seeger, M. , & Peters, J. . (2010). Real-time local gp model learning. Studies in Computational Intelligence,264(2010), 193-207.

[8] Schreiter, J. , Englert, P. , Nguyen-Tuong, D. , & Toussaint, M. . (2015). Sparse gaussian process regression for compliant, real-time robot control. Proceedings. IEEE International Conference on Robotics and Automation, 2015, 2586-2591.

[9] Kou, P. , & Gao, F. . (2014). Sparse gaussian process regression model based on?1/2regularization. Applied Intelligence, 40(4), 669-681.

[10] QuiñoneroCandela, J, Rasmussen, C. , Williams, C. , Chapelle, O. , Decoste, D. , & Weston, J. . (2007). Approximation Methods for Gaussian Process Regression. Large-Scale Kernel Machines. MIT Press.
[11] Ranganathan, A. , Yang, M. H. , & Ho, J. . (2011). Online sparse gaussian process regression and its applications. IEEE Transactions on Image Processing, 20(2), 391-404.
[12] Kou, P. , Gao, F. , & Guan, X. . (2013). Sparse online warped gaussian process for wind power probabilistic forecasting. Applied Energy, 108, 410-428.