Sparse Gaussian Process for Spatial Function Estimation with Mobile Sensor Networks

Bowen Lu, Dongbing Gu, Huosheng Hu and Klaus McDonald-Marier

***Abstract*— Gaussian process (GP) is well researched and used in machine learning field. Comparing with artificial neural network (ANN) and support vector regression (SVR), it provides additional covariance information for regression results. By exploiting this feature, an uncertainty based locational optimisation strategy combining with an entropy based data selection method for mobile sensor networks is presented in this paper. Centroidal Voronoi tessellation (CVT) is used as a locational optimisation framework and Informative Vector Machine (IVM) is applied for data selection. Simulations with different locational optimisation criteria are conducted and the results are given, which proved the effectiveness of presented strategy.**

I. Introduction

Wireless sensor networks (WSNs) are employed in various research fields which require to obtain sample data from a large scale of environment, such as forestry, meteorology, oceanography, etc [1][2]. For pollution monitoring, WSNs have been used for air and ocean environments in [3] and [4]. A sensor network is required to model the map of a spatial function in the environment. For a mobile sensor network, an effective locational optimisation strategy plays an important role on modelling performance.

Recently, kernel based regression methods are widely researched in machine learning field, including support vector machine (SVM) [5], relevance vector machine (RVM) [6], Kriging [7], and Gaussian process (GP) [8], etc. Some researchers found that these methods could convert from one to another under certain configurations [9]. Comparing among these kernel methods, the GP / Kriging gives additional uncertainty information for modelled distribution, which provides a criterion for locational optimisation.

Combining with the regression methods, various strategies are applied for locational optimisation in mobile sensor networks. Flocking algorithm was used with the spatial-temporal GP in [10]. Centroidal Voronoi tessellation (CVT) is another famous strategy introduced by Jorge Cortes from the view of computational geometry [11][12]. It was used with artificial neural network in [13].

Informative vector machine (IVM) is an entropy based method for selecting data from large number of samples, and was first presented by Neil Lawrence et al. in [14], [15], [16] and [17]. Their research shows that the IVM has a similar performance in computing speed and modelling accuracy with SVM.

The authors are with School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, UK. blv@esssex.ac.uk,dgu@essex.ac.uk, hhu@essex.ac.uk

In this paper, a framework of mobile sensor network for estimating a latent spatial function is given. More specifically, the CVT is used for optimising the sensor deployment; the GP estimates a latent spatial function and provides the uncertainty information for the model. In the CVT locational optimisation process, a combination of the mean and the covariance information is applied. Its performance is compared with using each of them alone. Potentially, a full data set is an ideal scenario for the modelling. However, using a full data set makes the computation of the GP model intractable. Therefore, the IVM is introduced to select a sub set with least information lost. The following sections are organised as below: section II-A gives a brief introduction to modelling the GP with a mobile sensor network, section II-B introduces the CVT with optimisation criteria configuration, and section II-C illustrates the principle of the IVM with data selection. Simulation results and analysis are given in section III.

II. Modelling Spatial Function with Wireless Sensor Networks

A. GP Approach to Estimating the Latent Function

To estimate a latent spatial function $f(x)$ in a $2D$ convex environment $\mathbb{Q}$, a mobile sensor network with N sensors is deployed. The $2D$ coordinates of sensor $i \in N$ are denoted by $x_i \in \mathbb{R}^2$ and its observation is $y_i \in \mathbb{R}$. By collecting data from the sensor network, a training data set $\mathcal{D} = [X, y]$ is constructed, where $X := [x_1, x_2, \ldots, x_N]^T$ and $y := [y_1, y_2, \ldots, y_N]^T$. With introducing a Gaussian noise ε_i to each sensor i, an observation model is given in eq. (1):

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. Corresponding to Bayesian framework with the latent function $f := [f(x_1), f(x_2), \ldots, f(x_N)]^T$, the observation y is the likelihood and its distribution can be illustrated as eq. (2):

$$p(y|f) \sim \mathcal{N}(y; f, \sigma_n^2 I) \tag{2}$$

where I is an identity matrix. The prior knowledge in the GP is defined by a kernel function $K(x_i, x_j)$ and here is modelled by:

$$K(x_i, x_j) = \sigma_f^2 \exp\left\{-\frac{\|x_i - x_j\|}{2l^2}\right\} \tag{3}$$

σ_f and l are two hyper-parameters, which can be modified online for controlling the amplitude and length scale of $K(x_i, x_j)$. The prior distribution of the latent function is given in eq. (4):

$$p(f|X) \sim \mathcal{N}(f; 0, K) \tag{4}$$

Let a test point be $\mathcal{D}_* = \{x_*, f_*\}$. A joint distribution between the observation y and the GP prediction f_* is obtained in eq. (5):

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K + \sigma_n^2 I & K_* \\ K_*^T & K_{**} \end{bmatrix}\right) \tag{5}$$

where x_* is an arbitrary test point in the environment $\mathbb{Q}$, and $f_* = f(x_*)$ denotes the predicted latent function value at x_*. In addition, K, K_* and K_{**} are shorthand notations, which denote $K(X, X)$, (X, x_*) and $K(x_*, x_*)$ respectively. The conditional distribution of f_* is obtained by conditioning eq. (5):

$$\begin{aligned} p(f_*|f, X, x_*) &\sim \mathcal{N}(\mu_*, \sigma_*) \\ \mu_* &= K_*^T[K + \sigma_n^2 I]^{-1} y \\ \sigma_* &= K_{**} - K_*^T[K + \sigma_n^2 I]^{-1} K_* \end{aligned} \tag{6}$$

μ_* and σ_* from eq. (6) are the estimated latent spatial function mean values and uncertainty at x_*. K, K_* and K_{**} are controlled by the hyper-parameters σ_f and l. In order to get the optimal values for σ_f and l, the maximum log marginal likelihood is applied as eq. (7):

$$\max_{\sigma_f, l} \{\log p(y|X)\} \tag{7}$$

where

$$\begin{aligned} &\log p(y|X) = \\ &-\frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{N}{2} \log(2\pi) \end{aligned} \tag{8}$$

B. A Heuristic potential function for the CVT Locational Optimisation

Centroidal Voronoi tessellation (CVT) is a gradient based locational optimisation method. To the CVT method for the sensor network in this paper, an arbitrary point from the environment $\mathbb{Q}$ is denoted by $\hat{x}$, hence the Voronoi cell of sensor i is defined as:

$$V_i = \{\hat{x} \in \mathbb{Q} | \ \|\hat{x} - x_i\| \leqslant \|\hat{x} - x_j\|, \forall i \neq j\} \tag{9}$$

Before applying the CVT to the sensor network, a potential function needs to be defined. The latent function model generated from the GP contains two parts, mean μ^* and covariance σ^*. Some potential functions are available for various situations: using mean component μ^* provides the smooth motion control result and an accurate local model, however it may be trapped at local mean maxima; using covariance component σ^* makes the sensor network cover as much area as possible, but it has lower accuracy on local details if the length scale is not large enough. Therefore, its modelling performance may end with a larger variance.

In balancing between them, we construct our potential function with the dot product of two components as equation (10):

$$f_p = \mu_{norm}^* \cdot \sigma_{norm}^* \tag{10}$$

where μ_{norm}^* and $\sigma_{norm}^* \in [0, 1]$ are normalised results from μ^* and σ^* respectively. With this combination form, the sensor network can escape from local mean maxima.

For each Voronoi cell V_i, its corresponding mass centre C_{V_i} is its optimal location:

$$\begin{aligned} M_{V_i} &= \int_{V_i} f_p(\hat{x}) d\hat{x} \\ L_{V_i} &= \int_{V_i} \hat{x} f_p(\hat{x}) d\hat{x} \\ C_{V_i} &= \frac{L_{V_i}}{M_{V_i}} \end{aligned}$$

C. Data selection: IVM

A full data set is constructed as $\mathcal{D}_f = \{\mathcal{D}_1, \ldots, \mathcal{D}_h\}$, where h indicates the length of $\mathcal{D}_f$. Normally, the criterion of selecting h depends on the changing speed of the latent spatial function. In balancing between the modelling accuracy and computing speed, an entropy based data selection method, the IVM [18] is introduced to select active data points from $\mathcal{D}_f$.

According to the principle of the IVM [18], it is necessary to update the posterior of the GP in a sequential mechanism. To achieve this goal, the IVM creates two index sets, active set I and inactive set J. $I = \emptyset$ and $J = \{1, \ldots, h \times N\}$ are defined as their initial. The index of data points is selected one after another from set J to I. An update form for computing the GP posterior estimation $\hat{p}_i$ with μ_i and σ_i is given (more details see [18]) as eq. (11), (12) and (13):

$$\hat{p}_i \sim \mathcal{N}(f; \mu_i, \sigma_i) \tag{11}$$

$$\mu_i = \mu_{i-1} + g_{n_i} \sigma_{i-1} e_{n_i} \tag{12}$$

$$\sigma_i = \sigma_{i-1} + (g_{n_i}^2 - 2G_{n_i}) \sigma_{i-1} e_{n_i} e_{n_i}^T \sigma_{i-1} \tag{13}$$

where

$$g_{n_i} = \frac{y_{n_i} - \mu_{i-1,n_i}}{\sigma_n^2 I + \sigma_{i-1,n_i}} \tag{14}$$

$$G_{n_i} = \frac{1}{2}\left(g_{n_i}^2 - \frac{1}{\sigma_n^2 I + \sigma_{i-1,n_i}}\right) \tag{15}$$

n_i denotes the data point indices in set J, and e_{n_i} is a unit vector choosing the n_ith element. μ_{i-1,n_i} and σ_{i-1,n_i} are the n_ith element of μ_{i-1} and the n_ith diagonal element of

σ_{i-1} respectively. With eq. (12), (13), the IVM can filter and select data points to active set I while the GP updating the posterior $\hat{p}_i$ in sequential. The criterion of the IVM selection is the change of information entropy, and is defined as eq. (16) after the ith data point is included.

$$H_i = H(\hat{p}_i) := -\int \hat{p}_i(f) \log \hat{p}_i(f) df \tag{16}$$

$\hat{p}_i$ is a Gaussian distribution, hence

$$H_i = \frac{i}{2}\log(2\pi e) + \frac{1}{2}\log|\sigma_i| \tag{17}$$

The entropy change after the ith data point is selected into active set I

$$\begin{aligned}\Delta H_{i,n_i} &= H_i - H_{i-1} \\ &= \underbrace{\frac{1}{2}\log(2\pi e)}_{\text{constant}} + \frac{1}{2}\log|\sigma_i \sigma_{i-1}^{-1}|\end{aligned} \tag{18}$$

It should be noticed that $\Delta H_{i,n_i}$ is negative when the entropy is reducing. Therefore, the data points with the smallest $\Delta H_{i,n_i}$ values are selected by the IVM.

III. Simulation Results

To simulate a $2D$ environment $\mathbb{Q}$, an 1×1 rectangle area is chosen. A sensor network with $N = 15$ is randomly deployed at a corner of $\mathbb{Q}$ ($x_i \in [0, 0.2]$). The sensor noise level $\sigma_n = 0.01$, the hyper-parameters $\sigma_f = 0.01$ and $l = 0.01$ are initialised. The full data set length $h = 10$ and 15 active data points are set for the IVM data selection. A latent spatial function f and its mesh graph are given in eq. (19) and shown in figure 1. An 100×100 evenly distributed grid mesh is configured for the environment $\mathbb{Q}$. The estimated latent spatial function model f_* is illustrated by two 100×100 matrices, μ^* and σ^*. Each value from the two matrices indicates the estimated mean value and the uncertainty level at a particular grid point respectively.

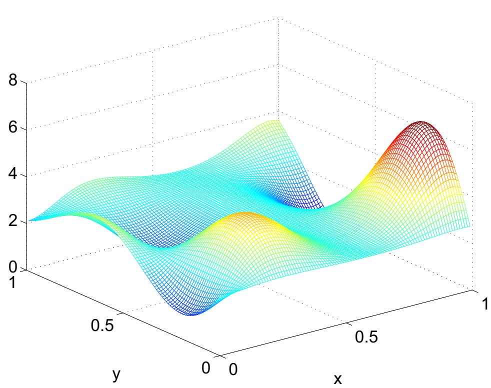

Fig. 1. Latent function in environment

$$f = 1.9\{1.35 + e^{x-y}\sin[13(x-0.6)^2]\sin(7y)\} \tag{19}$$

To illustrate the effectiveness of presented method, the simulation results are organised in three comparison groups and statistic data are collected from 100 tests. For the first group, the accuracy of estimated model from the IVM and the random selection are compared in figure 2. It can be observed that the IVM selection method provides a faster converging speed, smaller static error and standard deviation. In the next two groups, the IVM is employed as a data selection method to keep a fair comparison.

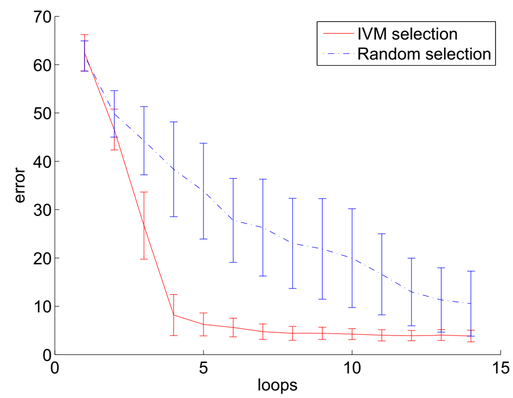

Fig. 2. Comparison between random selection and the IVM selection

The second group compares the estimation performances between two potential functions, $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$ and $f_p = \mu^*_{norm}$. As analysed in section II-B, using only μ^*_{norm} produces a smooth converging curve, and the error variance is quite stable. However it suffers from the local minima problem while the number of sensor nodes is insufficient to cover the whole environment with the learnt GP kernel length scale l. The model accuracy with the dot product form $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$ converges faster and stabler in the whole modelling process (smaller standard deviation).

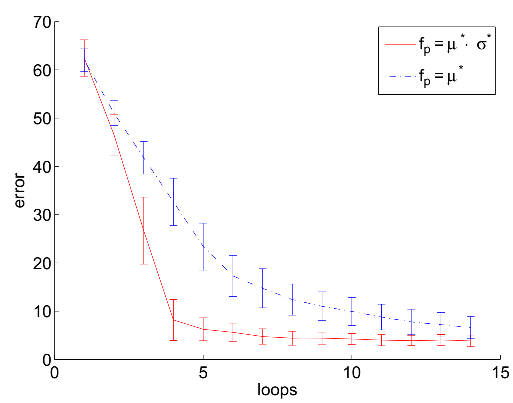

Fig. 3. Comparison between $f_p = \mu^*_{norm}$ and $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$

$f_p = \sigma^*_{norm}$ and $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$ are compared in the last group. It can be found that the accuracy from the uncertainty driven modelling process converges faster in the beginning, and keeps quite close to the performance with the combined potential function. Although these two potential functions give similar static errors to the true latent function f, the standard deviation from $f_p = \sigma^*_{norm}$ is much larger than the combined one.

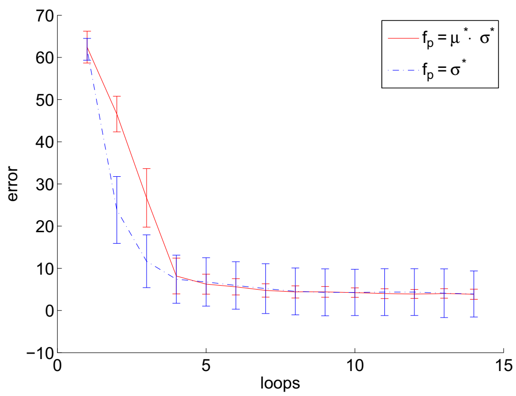

Fig. 4. Comparison between $f_p = \sigma^*_{norm}$ and $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$

Figure 5 shows one of the modelling process with $f_p = \mu^*_{norm} \cdot \sigma^*_{norm}$ and the IVM selection method. In figure 5, the estimated latent function is shown in top-left panel; Top-right panel illustrates the sensor deployment with red circles and selected data points with blue crosses. The hyper-parameter learning for the kernel length scale l and the amplitude σ_f is given in bottom-left and bottom-right panels respectively.

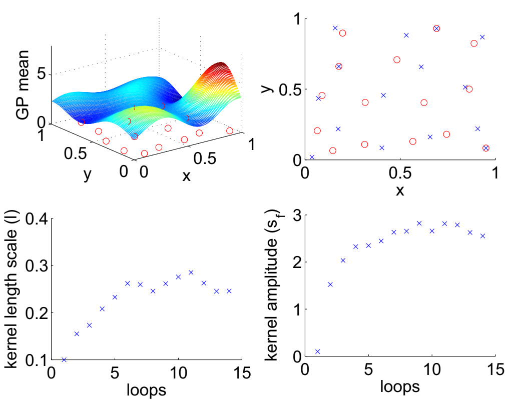

Fig. 5. Sensor deployment and hyper-parameters learning

IV. Conclusion

A dot product combination form with the GP mean and covariance information is used for the locational optimisation criterion in the CVT framework. With the comparison in section III, our presented potential function (eq. (10)) provides better performance in terms of converge speed and static error.

Introducing a data selection method, the IVM, a full data set is implemented without adding significant extra computational burden to the GP. The effectiveness of the IVM in data selection is proved by comparing with a random data selection process.

In the next step, a dynamic potential function could be studied to provide the flexibility to the sensor network with time variant or dynamic latent spatial function. Then the sensor network is able to change its locational optimisation criteria according to the real time information.

Acknowledgement: This research has been financially supported by EPSRC Global Engagements grant EP/K004638/1.

References

[1] N. Leonard, D. Paley, F. Lekien, R. Sepulchre, D. Fratantoni, and R. Davis, "Collective motion, sensor networks, and ocean sampling," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 48–74, 2007.

[2] C. Corrigan, G. Roberts, M. Ramana, D. Kim, V. Ramanathan *et al.*, "Capturing vertical profiles of aerosols and black carbon over the indian ocean using autonomous unmanned aerial vehicles," *Atmospheric Chemistry and Physics Discussions*, vol. 7, no. 4, pp. 11 429–11 463, 2007.

[3] W. Tsujita, H. Ishida, and T. Moriizumi, "Dynamic gas sensor network for air pollution monitoring and its auto-calibration," in *Sensors, 2004. Proceedings of IEEE*, oct. 2004, pp. 56 – 59 vol.1.

[4] A. Khan and L. Jenkins, "Undersea wireless sensor network for ocean pollution prevention," in *Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference on*, jan. 2008, pp. 2 –8.

[5] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

[6] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[7] M. Oliver and R. Webster, "Kriging: a method of interpolation for geographical information systems," *International Journal of Geographical Information System*, vol. 4, no. 3, pp. 313–332, 1990.

[8] C. Rasmussen, "Gaussian processes in machine learning," *Advanced Lectures on Machine Learning*, pp. 63–71, 2004.

[9] W. Chu, S. Keerthi, and C. Ong, "Bayesian support vector regression using a unified loss function," *Neural Networks, IEEE Transactions on*, vol. 15, no. 1, pp. 29–44, 2004.

[10] J. Choi, J. Lee, and S. Oh, "Swarm intelligence for achieving the global maximum using spatio-temporal Gaussian processes," in *Proc. of the American Control Conference*, Seattle, Washington, 2008.

[11] J. Cortes, S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Trans. on Robotics and Autonamous*, vol. 20, no. 2, pp. 243–255, 2004.

[12] L. C. A. Pimenta, V. Kumar, R. C. Mesquita, and G. A. S. Pereira, "Sensing and coverage for a network of heterogeneous robots," in *Proc. of the IEEE Conf. on Decision and Control*, Cancun, Mexica, 2008.

[13] M. Schwager, D. Rus, and J. Slotine, "Decentralized, adaptive coverage control for networked robots," *Int. J. of Robotics Research*, vol. 28, no. 3, pp. 357–375, 2009.

[14] N. Lawrence, M. Seeger, and R. Herbrich, "Fast sparse gaussian process methods: The informative vector machine," *Advances in neural information processing systems*, vol. 15, pp. 609–616, 2002.

[15] N. Lawrence and M. Jordan, "Gaussian processes and the null-category noise model," *Semi-Supervised Learning*, pp. 137–150, 2006.

[16] ——, "Semi-supervised learning via gaussian processes," *Advances in neural information processing systems*, vol. 17, pp. 753–760, 2005.

[17] N. Lawrence, J. Platt, and M. Jordan, "Extensions of the informative vector machine," *Deterministic and Statistical Methods in Machine Learning*, pp. 56–87, 2005.

[18] N. Lawrence, M. Seeger, and R. Herbrich, "The informative vector machine: A practical probabilistic alternative to the support vector machine," *Dept. Computer Science, University of Sheffield, Tech. Rep*, 2005.