Improved Building Detection by Gaussian Processes Classification via Feature Space Rescale and Spectral Kernel Selection

Hang Zhou
Dept Elec. & Comp. Syst. Eng.
Monash University, Clayton, VIC 3800, Australia
hang.zhou@eng.monash.edu.au

David Suter
Dept Elec. & Comp. Syst. Eng.
Monash University, Clayton, VIC 3800, Australia
d.suter@eng.monash.edu.au

Abstract

We use spectral analysis to facilitate Gaussian processes (GP) classification. Our solution provides two improvements: scaling of the data to achieve a more isotropic nature, as well as a method to choose the kernel to match certain data characteristics. Given the dataset, from the Fourier transform of the training data we compare the frequency domain features of each dimension to estimate a rescaling (towards making the data isotropic). Also, the spectrum of the training data is compared with several candidate kernel spectrums. From this comparison the best matching kernel is chosen. In these ways, the training data matches better the GP classification kernel function (and hence the underlying assumed correlation characteristics), resulting in a better GP classification result. Test results on both non image and image data show the efficiency and effectiveness of our approach.

1 Introduction

We aim to develop an efficient way of improving the performance of building detection (segmentation) based on Gaussian processes (GP) classification.

The GP classification process models the posterior directly, thus relaxing the strong assumption of conditional independence of the observed data (generally used in a generative model).

The GP prior is represented by the kernel function which characterizes correlations between points in the training data (which is a sample process). The kernel function hyperparameters can be learned from the training data.

We found two issues exist in GP classification. The GP classification performance with the same kernel function (isotropic) on different data (usually anisotropic) can vary significantly. The GP classification results are also different when applying different kernel functions to the same data. Essentially, the two issues are aspects of the problem of how the isotropic GP kernel prior can be matched with data having varied characteristics.

Obviously, an anisotropic kernel function can be used to cope with the anisotropic data, at the cost of greatly increasing the complexity.

Our approach improves the GP classification performance by means of better matching the GP kernel with the data from two aspects: First, minimize the training data spectrum differences, between dimensions (feature data), by rescaling the training data input feature space on each dimension (so as to make a better match with the isotropic GP kernel function). The parameter of rescaling is calculated adaptively, based on the analysis on frequency domain, where the data characteristics are more distinguishable.

Secondly, estimate the spectrum of the rescaled data on each dimension, which is then compared with several kernel spectrums from which the most matched kernel is chosen.

Existing approaches for building detection include Kumar and Hebert's approaches using Multiscale Random Field (MSRF) and Discriminative Random Field (DRF) models whose results will be compared to ours. Lin and Nevatia [1] use rooftop and aerial images but these have different characteristics to ground level building views (the focus of this paper).

Alternative solutions for tackling data anisotropy and non-stationarity include data partitioning [2] which deals with the special case of sharp changing data. Non-stationary kernel functions [3] and mixtures of stationary GP [4] have also been used.

Specifically focusing on the anisotropy problem, Schmidt and O'Hagan [5], applied an interpolation/ deformation: which maps the original space to a new isotropic one. The approach employs Monte Carlo Markov Chain (MCMC) methods, making the solution involved and computationally costly. Snelson, Rasmussen and Ghahramani presented a warped GP [6] where the

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

transformations are applied to the observation (output) space and make the data better modelled by a GP.

As for kernel selection which is part of the model selection problem in GP classification, existing methods include Bayesian model selection and cross validation [7]. These methods work reasonably well for optimization of the kernel hyperparameters. When dealing with choice of the functional form kernel, these approaches would either be intractable or at least very difficult.

In our approach, similar to the scheme of Snelson et. al [6], instead of warping the output space, the input feature space is rescaled to be more isotropic but in a more simple way. We seek to address the kernel-data matching problem efficiently by analyzing the training data, as well as the kernel function, in the frequency domain. We then apply a scale transformation on the input feature data space followed by a comparison between kernel and training data spectrums for kernel selection. No costly computation is involved.

The paper is structured as follows. GP spectral analysis on kernel functions is introduced in Section 2. A description of the kernel-data matching algorithm is given in Section 3. In Section 4, experiment details and results are presented. Section 5 provides the main conclusions of the work.

2 Gaussian processes spectral analysis on kernel functions

A GP is fully specified by its mean function m(x) and kernel function k(x, x') , expressed as:

$$f \sim GP(m,k) \tag{2.1}$$

In the building detection application, a binary GP classifier is needed to discriminate between man-made structure and non-structure assuming the dataset is $D=(X,y)$, where X are input training data features and y the class labels -1/+1.

GP binary classification is done by first calculating the distribution over the latent function, then the output of regression is 'squashed' through a sigmoid transformation to guarantee the valid probabilistic value within the range of [0,1].

The GP kernel is the crucial part of GP classification as it incorporates the prior smoothness assumption. The kernel functions studied in this paper include:

1) Radial Basis Function (RBF), also called as Squared Exponential (SE) function or Gaussian function

$$k_{RBF}(r)=\exp(-\frac{r^2}{2l^2}) \tag{2.2}$$

where $r=x-x'$, x and x' are input pairs, l is the characteristic length-scale.

2) Matern class of covariance functions

$$k_{\nu=5/2}(r)=(1+\frac{\sqrt{5}r}{l}+\frac{5r^2}{3l^2})\exp(-\frac{\sqrt{5}r}{l}) \tag{2.3}$$

$$k_{\nu=3/2}(r)=(1+\frac{\sqrt{3}r}{l})\exp(-\frac{\sqrt{3}r}{l}) \tag{2.4}$$

where r and l are the same as in equation (2.2).

We denote the kernel functions in equation (2.2), (2.3) and (2.4) as RBF, M52 and M32 respectively. Their corresponding spectrums in D dimension are represented as follow:

$$S_{RBF}(s)=\left(2\pi l^2\right)^{D/2}\exp(-2\pi^2 l^2 s^2) \tag{2.5}$$

$$S_{M52}(s)=\frac{2^D\pi^{D/2}\Gamma(2.5+D/2)}{\Gamma(2.5)l^5}(\frac{5}{l^2}+4\pi^2 s^2)^{-(2.5+D/2)} \tag{2.6}$$

$$S_{M32}(s)=\frac{2^D\pi^{D/2}\Gamma(1.5+D/2)}{\Gamma(1.5)l^3}(\frac{3}{l^2}+4\pi^2 s^2)^{-(1.5+D/2)} \tag{2.7}$$

From equation (2.5), (2.6) and (2.7):

$$S_{RBF}(s)\propto \exp^{-s^2} \tag{2.8}$$

$$S_{M52}(s)\propto s^{-(5+D)} \tag{2.9}$$

$$S_{M32}(s)\propto s^{-(3+D)} \tag{2.10}$$

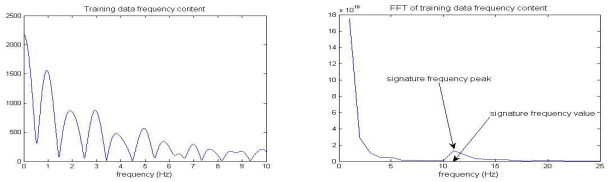
Further details of kernels can be found in [7].

3 The kernel-data matching algorithm

The kernel-data matching algorithm is a two step procedure: First, training data is rescaled so that the spectrum difference between dimensions is minimized for the purpose of better fit to the isotropic kernels. Then the spectrum is estimated from this rescaled data to match with the spectrum of one of a group of kernels (see above) from which the best kernel is chosen.

3.1 Data rescale

The GP classification with isotropic kernel function works better on 'isotropic' data, i.e. data with homogeneous properties on each dimension. We focus on a property which we call 'signature frequency', and, intuitively, this describes the fluctuation of training data frequency content in Figure 1(a). This is further explained below.



(a) (b)

Figure 1. Signature frequency. (a) Training data frequency content (on one dimension). (b) FFT of (a).

In order to analyse the training data on each dimension individually and efficiently, the high dimensionality of the unevenly spaced training input is reduced by projecting to each dimension where the nonuniform discrete Fourier transform (NDFT) [8] is calculated. Thereby, data rescaling is carried out by simply rescaling training input respectively on each dimension so as to make the 'signature frequency' of the frequency content on different dimensions consistent with each other.

The 'signature frequency' on each dimension is estimated in the following way:

1) Project training data onto to each dimension.
2) Calculate the NDFT of the projected training data on each dimension (which is called training data frequency content hereafter).
3) Further calculate the FFT of the training data frequency content on each dimension to capture the frequency fluctuation.
4) Smooth (using any common method like kernel smoothing) the FFT coefficients to ignore small fluctuation.
5) If only one peak exist on the FFT coefficients, choose this peak and record the corresponding frequency value as its 'signature frequency'.
6) If more than one peak exist among the FFT coefficients as in Figure 1(b), choose the one within high frequency range, which is empirically defined as greater than 4 Hz. Record the related frequency value as 'signature frequency'.
7) Set the maximum 'signature frequency' value over all dimensions as the 'target frequency'.
8) Rescale index = target freq. / signature freq. (3.1)
9) Multiply the input of the training data with "rescale index".

Thus, rescaling is implemented in space domain, making the signature frequency close to the same target frequency value on each dimension.

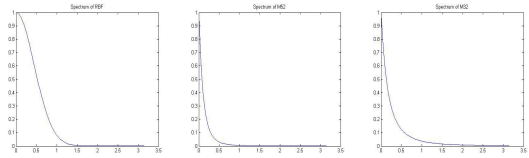
3.2 Kernel selection

The assumption in GP classification is that the training data is a sample drawn from the GP process specified by a particular kernel function. It is further assumed, in engineering applications, that the GP is an ergodic process: which means that the time average and the ensemble average are the same. In this way, the spectrum of the GP can be estimated from one of its derived observed sample [9] (which is the training data in our application) rather than many samples over a long period of time. From the rescaled data obtained in Section 3.1, the spectrum is further estimated to choose the best kernel.

It should be noted that (different from the calculation in Section 3.1 which is to obtain the frequency content of the training data itself) what is to be estimated here is the spectrum of the underlying GP that derives the training data as one manifestation of the process.

By estimating the kernel spectrum from the given training data, and matching that with one of the candidate kernel spectrums, the best kernel can be chosen for the training data.

Since the rescaled data obtained in Section 3.1 are still unevenly distributed (multiplying a coefficient does not change the nonuniformity of the data distribution), conventional spectrum estimation methods which are based on equally spaced data cannot be used directly. Therefore, an IFFT is implemented on the rescaled data's NDFT result (which can be deduced from the NDFT result in Section 3.1 – see step 1 below). The IFFT is then used to produce equivalent (but equally spaced) version of the rescaled nonuniform data. The underlying GP spectrum can then be estimated on this equivalent evenly spaced data.



(a) (b) (c)

Figure 2. Reference spectrums. (a) RBF. (b) M52. (c) M32.

Among the candidate kernels, the one with its spectrum most correlated with the underlying GP spectrum of the training data, is chosen to be the best kernel. Figure 2 shows the candidate kernel spectrums w.r.t. equation (2.8) ~ (2.10) which are used as reference spectrums for the underlying GP spectrum to compare with. Parameters of the reference spectrums in Figure 2 are adjusted for the function values to converge at 1.5 Hz so as to keep consistency between spectrums.

The detailed algorithm is as follows:

1) Calculate the NDFT of the rescaled data (which can be done by scaling, by the reciprocal of the rescale index, the NDFT result of Section 3.1).
2) Calculate IFFT of the NDFT result from 1) to get an equivalent equally spaced data of the original rescaled data.
3) Estimate the underlying GP spectrum on each dimension using the eigenvector method - which is one of the many existing spectrum estimation methods

available in Matlab ("peig" function in Signal Processing Toolbox).

4) Calculate the correlation coefficients of the estimated underlying spectrum with the reference kernel spectrums (illustrated in Figure 2).
5) Compare the correlation coefficients of each dimension across all kernels.
6) Choose the kernel with the property that its correlation coefficient is the highest across the maximum number of dimensions.

4 Experiments and results

4.1 Non image data

Our approach has been tested on non-image data besides image data.

The 8 non-image data files are randomly taken from 'The UCI Repository of Machine Learning Databases and Domain Theories'.

Figure 3 shows the rescaling results on file 'Monks-3'.

Each dimension of the rescaled data is more homogeneous in the frequency domain w.r.t. the 'fluctuation': as a result of scale transformation with parameters shown in Table 1.

	Signature frequency	Rescale index= max(signature frequency) /signature frequency
Dim 1	20	1
Dim 2	10	2
Dim 3	10	2
Dim 4	20	1
Dim 5	10	2
Dim 6	20	1
Dim 7	20	1

Table 1. Signature frequency and 'rescale index' values of 'Monks-3' on each dimension.

	RBF	M52	M32
Dim 1	0.9436	0.9026	0.9822
Dim 2	0.8836	0.8488	0.9211
Dim 3	0.8926	0.8271	0.9098
Dim 4	0.9436	0.9026	0.9822
Dim 5	0.4622	0.7677	0.7064
Dim 6	0.4622	0.7677	0.7064
Dim 7	0.4622	0.7677	0.7064
Max count	0	3	4
Improved GP DR	0.9035	0.9079	0.9167
Improved GP FP	4	4	3
Standard GP DR	0.8860	0.8947	0.8991
Standard GP FP	8	8	7

Table 2. **C**orrelation coefficients between data spectrums and kernel spectrums as well as the classification performance on different kernels (DR denotes detection rate and FP denotes false positives).

Kernel selection is then implemented based on the rescaled data using the algorithm described in Section 3.2 with details listed in Table 2. The first 7 rows show the correlation coefficients between the spectrum on each dimension and the candidate kernel function spectrums. The kernel with the most maximum coefficients is chosen. In this case, M32 is chosen (with 4 maximum coefficients).

The results of our improved GP classification, as listed in Table 2, compared with standard GP classification results, clearly shows that the rescaling improves GP classification performance on all kernels. Moreover, our kernel selection clearly chooses the best kernel (M32 in all of these cases).

Test results on more UCI data files are provided in Figure 4: note that the performance is enhanced resulting in either an increase on detection rate or a drop on false positives (or both). This shows that our algorithm works well on varied datasets.

4.2 Image data

The proposed approach was trained and tested using the Corel images that Kumar [10] used[1]. All images are cut to the size of 256x256 and the training images are divided into non-overlapping 16x16 pixels blocks which are labelled as one of the two classes, i.e. building or non-building blocks.

A 14 component feature vector is computed at each 16×16 block. These features are derived from "orientograms" which are the histograms of gradient orientations in a region weighted by gradient magnitudes. They are designed to capture the lines and edges patterns in man-made structures [10] [11].

We used a training set of 28 Corel images, containing 714 structured blocks and 2893 non-structured blocks. Testing is implemented on 70 Corel images.

We run Lawrence's program [12] [2] for GP classification.

Figure 5 shows a more similar frequency content in each dimension, after the data has been rescaled. Especially, there is a significant change on dimension 4 to 6 (a large scaling in the signal domain, with the parameters shown in Table 3, resulting a shrink in frequency domain).The frequency content is more concentrated in the lower frequency, which is more consistent with the kernel function spectrum.

Results of (subsequent) kernel selection are presented in Table 4 - where the maximum correlation on each row "vote" for a kernel type and the kernel with the largest "vote"(Max count), i.e. M32 is chosen. From the results in Table 5, it can be seen for our improved GP

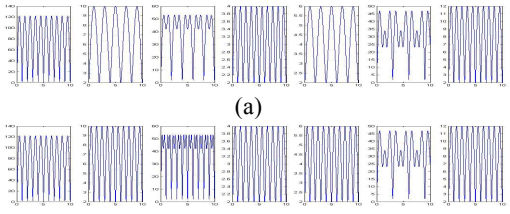
[1] http://www.cs.cmu.edu/~skumar/manMadeData.tar

[2] http://www.cs.man.ac.uk/~neill/ivm/downloadFiles/

classification, with training data properly rescaled, and the best kernel (which is M32) chosen; the performance (i.e. detection rate and false positives), are clearly better than either the standard GP results or Kumar's [10, 13] MSRF and DRF results.

It should also be noticed that under our improved GP classification framework, M32 kernel performs better than RBF. However, this contrasts with standard GP classification where the RBF has a better result.

The tables report average performances. Of course this is the result of higher detection rates and lower false positives obtained on many individual images: some sample images are shown in Figure 6. Improved GP classification result images are compared with standard GP classification images as well as Kumar's [10, 13] image results (reproduced from his black and white format papers). One can see that our improved GP classification produces building labels (the squares) that tend to cover more building regions and have less false detections.



(a)

(b)

Figure 3. Frequency content of 'Monks-3' (a) Original frequency content on each dimension. (b) Rescaled frequency content on each dimension.

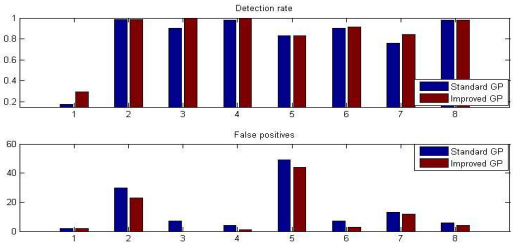
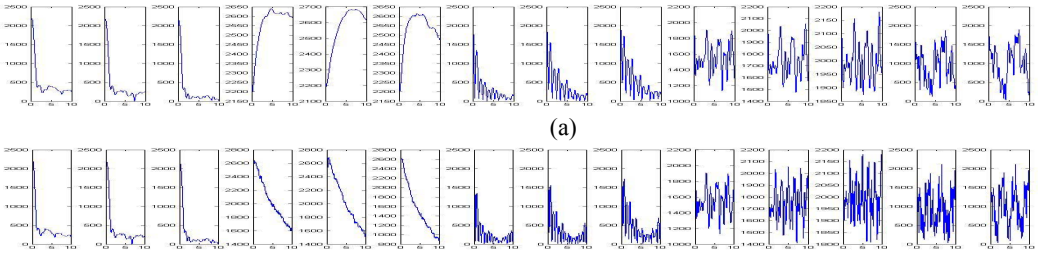


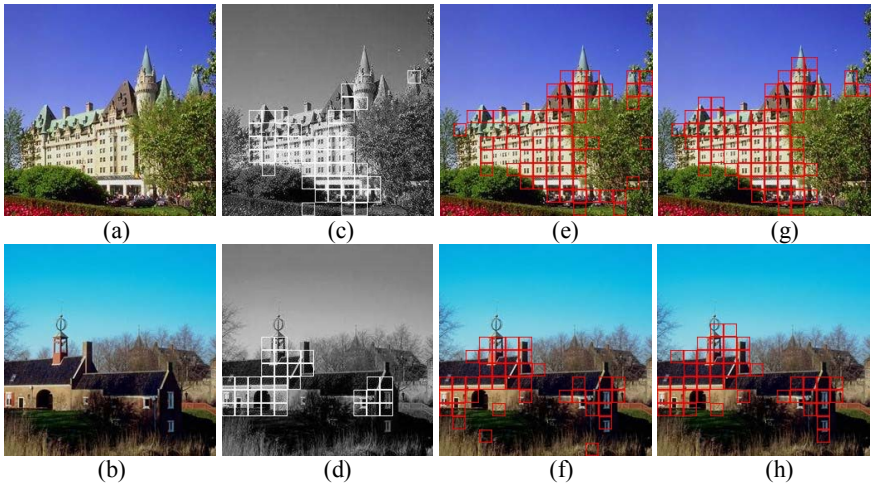
Figure 4. Performance comparison between standard GP and improved GP on UCI data.



(a)

(b)

Figure 5. Frequency content of Corel images. (a) Original frequency content on each dimension. (b) Rescaled frequency content on each dimension.



(a) (c) (e) (g)

(b) (d) (f) (h)

Figure 6. Classification results. (a)(b) Original images (c)(d) Kumar's results (e)(f) Standard GP results (g)(h) Improved GP results.

	Signature frequency	Rescale index= max(signature frequency) /signature frequency
Dim 1	10	1.3
Dim 2	12	1.08
Dim 3	13	1
Dim 4	1	13
Dim 5	1	13
Dim 6	1	13
Dim 7	10	1.3
Dim 8	10	1.3
Dim 9	10	1.3
Dim 10	12	1.08
Dim 11	7	1.86
Dim 12	7	1.86
Dim 13	6	2.17
Dim 14	6	2.17

Table 3. Signature frequency and rescale index values of Corel images on each dimension.

	RBF	M52	M32
Dim 1	0.9431	0.8985	0.9788
Dim 2	0.9515	0.8906	0.9768
Dim 3	0.9730	0.6986	0.8543
Dim 4	0.8443	0.9692	0.9887
Dim 5	0.8421	0.9631	0.9844
Dim 6	0.8690	0.9549	0.9888
Dim 7	0.5475	0.9440	0.8430
Dim 8	0.4483	0.7191	0.6790
Dim 9	0.3655	0.6115	0.5773
Dim 10	0.7879	0.9887	0.9776
Dim 11	0.9138	0.9264	0.9865
Dim 12	0.8812	0.9483	0.9890
Dim 13	0.9062	0.9424	0.9916
Dim 14	0.9548	0.8838	0.9714
Max count	1	4	9

Table 4. Correlation coefficients between the data spectrums and kernel spectrums.

	Detection rate	False positives
Kumar's MRSF	0.7213	1.46
Kumar's DRF	0.7050	1.37
Improved GP - M32	0.7660	1.23
Improved GP - RBF	0.7540	1.40
Standard GP - M32	0.7250	2.79
Standard GP - RBF	0.7370	2.06

Table 5. Performance comparison of Corel images classification with improved and standard GP results as well as Kumar's results.

5 Conclusions

We proposed an efficient, yet effective, way to improve the GP classification by exploiting spectral analysis: which is effective in revealing the underlying characteristics on both training data and kernel functions. The training data is better matched with the GP kernel function by rescaling to improve isotropy; as well as by kernel selection based on spectrum comparison. As a result, a better GP classification performance is achieved.

More sophisticated feature space adjustment other than scale transformation can be further investigated. The solution can also be extended to more kernel functions.

6 References

[1] C. Lin and R. Nevatia, "Building Detection and Description from a Single Intensity Image," *Computer Vision and Image Understanding,* vol. 72, pp. 101-121, 1998.

[2] H.-M. Kim, B. K. Mallick, and C. C. Holmes, "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," *Journal of American Statistical Association,* vol. 100, pp. 653-668, 2005.

[3] C. J. Paciorek and M. J. Schervish, "Nonstationary covariance functions for Gaussian process regression," in *Neural Information Processing Systems*, 2003.

[4] C. E. Rasmussen and Z. Ghahramani, "Infinite Mixtures of Gaussian Process Experts," *Advances in Neural Information Processing Systems,* vol. 14, 2002.

[5] A. M. Schmidt and A. O'Hagan, "Bayesian Inference for Nonstationary Spatial Covariance Structures via Spatial Deformations," *Journal of the Royal Statistical Society Series B,* vol. 65, pp. 743-758, 2003.

[6] E. Snelson, C. E. Rasmussen, and Z. Ghahramani, "Warped Gaussian Processes," in *Neural Information Processing Systems*, 2003.

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

[8] S. Bagchi and S. K. Mitra, *The Nonuniform Discrete Fourier Transform and Its Applications in Signal Processing*: Kluwer Academic Publishers, 1999.

[9] "Power Spectrum Estimation," in *Noise and Random Processes Lectures*: Centre for Image Science, Rochester Institute of Technology, 2002.

[10] S. Kumar and M. Hebert, "Man-Made Structure Detection in Natural Images using a Causal Multiscale Random Field," in *CVPR2003*, 2003, p. 119.

[11] C. Pantofaru, R. Unnikrishnan, and M. Hebert, "Toward Generating Labeled Maps from Color and Range Data for Robot Navigation," in *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2003.

[12] N. D. Lawrence, J. C. Platt, and M. I. Jordan, "Extensions of the Informative Vector Machine," in *Deterministic and Statistical Methods in Machine Learning*, 2004.

[13] S. Kumar and M. Hebert, "Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification," in *Internaltional Conference on Computer Vision (ICCV'03)*, 2003.