Integration of Gaussian Process and MRF for Hyperspectral Image Classification

Wentong Liao, Jun Tang, Bodo Rosenhahn, Micheal Ying Yang
Institute for Information Processing (TNT), Leibniz University Hannover
Appelstr. 9A, 30167 Hannover, Germany
Email: {liao, jtang, rosenhahn, yang}@tnt.uni-hannover.de

Abstract—**In this paper, we propose a framework GP-MRF, which combines Gaussian processes (GPs) and Markov random field (MRF) for accurate classification of hyperspectral remote sensing image (HSI) data. This method exploits the relationship among adjacent pixels and integrates it into spectral information to obtain spectral-spatial classification. This framework consists of two steps. Firstly, a GP classifier (GPC) yields pixelwise predictive probability for each class. Secondly, an MRF is applied to extract spatial contextual information in the label map achieved in the first step. Then the classification results are inferred from the spectral-spatial information. By means of MRF regularization an enhanced classification result has been obtained. The experiments are performed on three hyperspectral benchmark datasets. The results from the GPC are compared with those obtained by state-of-the-art classification approaches and demonstrate that, GP model is a competitive tool for classification of HSI in terms of accuracy. Furthermore, the experimental results indicate that our proposed method GP-MRF improves the classification accuracy of conventional GPC.**

I. Introduction

The abundant spectral information contained in hyperpspectral data enables the characterization, identification, and classification of the land-covers with improved accuracy and robustness. However, several critical problems are unevadable in classification of HSI, among which: 1) a great number of spectral bands and relatively a small number of labeled training samples, which poses the well-known Hughes phenomenon [1]; 2) the spatial variability of the spectral signature; 3) noisy environment; 4) The scene of different objects made by the same or similar material (e.g. the roofs of some buildings and the roads can be made by the same material, asphalt) makes it hard to distinguish different land-covers. Therefore, the contextual information is necessary for classification task of HSI.

In recent years, some state-of-the-art methods have been successfully applied in the remote sensing community to classification task, such as support vector machine (SVM) [2] and random forest (RF) [3]. In particular, the kernel-based methods represented by SVM have been proved as an excellent classification approach for HSI in terms of accuracy and robustness [2][4]. The kernel-based methods have the inherent virtues: 1) handling high dimensional input spaces efficiently; 2) dealing with noisy samples in a robust way; 3) work with a relatively low number of labeled training samples. These properties make them well-suited to tackle the classification problems of HSI. GPs are another representative of potentially promising kernel-based methods. They have been successfully applied to HSI classification and yielded comparable or even better performance than SVM in terms of accuracy [5]. Moreover, they provide truly probabilistic outputs with an explicit degree of prediction uncertainty. In contrast to non-probabilistic approaches, the probabilistic techniques have various advantages in practical recognition circumstances [6]. Furthermore, there exist algorithms for GP hyperparameter learning which are lacking in the SVM framework. Therefore, GP is more likely to yield better classification results. However, Bayesian GP methods have not received much attention from remote sensing image community.

In order to alleviate the aforementioned spatial problems, it is necessary to exploit spatial contextual information to enhance the classification accuracy that is only based on spectral information. Markov random fields (MRFs) are effective probabilistic models to integrate spatial correlation of neighbours in a label image into a decision rule [7]. The maximum a posterior (MAP) decision rule is typically used in this framework [8]. In the MRF model, we assume that the class distribution of each pixel depends on a certain degree on its adjacent pixels. This assumption is reasonable because of two practical reasons: 1) adjacent pixels have mixed spectral response on the center pixel, especially the pixels near the borders (spatial boundaries); 2) in a HSI over an urban or suburban region, each land-cover type mostly arises in form of a patch, lump or local region. In mostly pixelwise classification results of HSI we observe that, many scattered pixels are assigned different labels from its adjacent pixels, or a small plot among a big region is classified as another land-cover type. Such classification results are normally suspectable. By means of combination of spectral information with spatial contextual information to construct a new decision rule the classification results can be modify and the accuracy will be clearly enhanced.

In this paper, we present a GP- and MRF-based (GP-MRF) method for spectral-spatial classification of HSI. Firstly, a GP model is applied to obtain the label image of HSI and yield predictive probability of each pixel for each class. Secondly, spatial contextual information is extracted by MRF model based on the label map. Finally, the spectral information is integrated into spatial contextual information to construct a new decision rule and each pixel will be reclassified. The second and third steps will be repeated until the results satisfy a predefined criterion.

This paper is outlined as follows. Section II briefly reviews the formulation of GPC and MRF, and then discusses how to combine this two methods. Section III presents and discusses the experimental results. We conclude the paper in Section IV.

978-1-4799-6652-3/15/$31.00~ ©2015 IEEE

II. GP-MRF MODEL

A. GP model for classification

Given a training set $(\mathbf{X},\mathbf{Y}) = \{\mathbf{X}_n, Y_n\}_{n=1}^{N}$, where N is the number of labeled samples and Y_n is the corresponding class label that indicates the land-cover type. Each vector $\mathbf{X}_n \in R^d$ represents the spectral d bands of a pixel in a HSI. Our task is labeling a new test sample set $\mathrm{x} = \{\mathrm{x}_m\}_{m=1}^{M}$, where M is the number of test samples, by computing the probability $P(y|\mathbf{X},\mathbf{Y},\mathrm{x})$ belonging to a class. For simple illustrating the binary classification with target $y_i \in \{-1,+1\}$ is considered here. The binary classification is easily extended to multiple classification by using the one-against-all or one-against-one strategy.

GP models generate a discrete label y_i for a data point x_i via a continuous latent variable f_i [9]. A likelihood model $p(\mathbf{y}|\mathbf{f})$ characterizes the monotonic relationship between latent variable $\mathbf{f}$ and the probably observed annotation $\mathbf{y}$. Several forms of squashing functions are available for such likelihood model. In particular the logistic and probit function are the most popular. In this paper, the probit function is considered.

$$p(y_i = +1|f_i) = \varphi(y_i f_i) \tag{1}$$

where φ is the Gaussian cumulative distribution function with the form:

$$\varphi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})dx \tag{2}$$

To make a probability prediction for x an integrating over the latent variable f is executed as follows:

$$p(y_i = +1|\mathbf{X},\mathbf{Y},\mathrm{x}_i) = \int p(y_i|f_i)p(f_i|\mathbf{X},\mathbf{Y},\mathrm{x}_i)df_i \tag{3}$$

where $p(f_i|\mathbf{X},\mathbf{Y},\mathrm{x}_i)$ is the distribution of latent variable f_i corresponding to x_i. It can be obtained by integrating over $\mathbf{F} = (F_1,\ldots,F_n)$, which is the latent variable corresponding to training set $(\mathbf{X},\mathbf{Y})$:

$$p(f_i|\mathbf{X},\mathbf{Y},\mathrm{x}_i) = \int p(f_i|\mathbf{X},\mathbf{Y},\mathrm{x}_i,\mathbf{F})p(\mathbf{F}|\mathbf{X},\mathbf{Y})d\mathbf{F} \tag{4}$$

where $p(\mathbf{F}|\mathbf{X},\mathbf{Y}) = p(\mathbf{F}|\mathbf{Y})p(\mathbf{F}|\mathbf{X})\ /\ p(\mathbf{Y}|\mathbf{X})$ is the posterior over the latent variables. $p(\mathbf{Y}|\mathbf{X})$ is the marginal likelihood (evidence), $p(\mathbf{F}|\mathbf{X})$ is the GP prior over the latent function, which in GP model is a jointly zero mean Gaussian distribution and with the covariance given by the kernel K.

The non-Gaussian likelihood in Eq. (4) makes the integral analytically intractable. We have to resort to either analytical approximation of integrals or Monte Carlo methods. Two well known analytical approximation methods are very suitable for this task, namely the *Laplace* [10] and the *Expectation Propagation* (EP) [11]. They both approximate the non-Gaussian joint posterior as a Gaussian one. In this paper we adopt the *Laplace* method since its computation cost relative lower than EP with comparable accuracy. As introduced in [9] the mean and variance of f_i are obtained as follow:

$$\mu_i = \mathbf{k}(\mathrm{x}_i)^T\mathbf{K}^{-}\widetilde{\mathbf{F}} \tag{5}$$

$$\sigma^{\mathbf{2}}{}_i = \mathbf{k}(\mathrm{x}_i,\mathrm{x}_i) - \mathbf{k}(\mathrm{x}_i)^T(\mathbf{K}+\mathbf{W}^{-})\mathbf{k}(\mathrm{x}_i) \tag{6}$$

where $\mathbf{W} \triangleq -\nabla\nabla logp(\mathbf{Y}|\widetilde{\mathbf{F}})$ is diagonal. $\mathbf{K}$ denotes a $N-by-N$ covariance matrix between N training points. $\mathbf{k}(\mathrm{x}_i)$ is a covariance vector between N training points X and a test points x_i and $\mathbf{k}(\mathrm{x}_i,\mathrm{x}_i)$ is covariance for test point x_i and $\widetilde{\mathbf{F}} = \arg\max_{\mathbf{F}} p(\mathbf{F}|\mathbf{X},\mathbf{Y})$. Given the mean and variance of f_i, we compute the prediction probability in Eq. (3).

The covariance function is the crucial ingredient in GP predictor and its hyperparameters Θ crucially affect its performance. The Gaussian radial basis function (RBF) is one of the most widely used kernels since its robustness for different types of data and given as follow:

$$K_{RBF}(\mathrm{x}_i,\mathrm{x}_j) = \sigma^2 exp(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\ \|^2}{2l^2}) \tag{7}$$

$\Theta = [\sigma, l]$ is the hyperparameter set for RBF, of which l in the function is the characteristic lengthscale, which informally can be roughly considered as the distance you have to move in input space for the function value to become uncorrelated. The smaller l we choose, the more rapidly the function varies. In this case, all of the training points are more correctly classified. Moreover, if l varies with input dimensions (i.e. input bands), e.g. $l = [l_1,\ldots,l_d]$, there is another kernel called the Automatic Relevance Determination (ARD) which is derived form RBF:

$$K_{ARD}(\mathrm{x}_i,\mathrm{x}_j) = \sigma^2 exp(-\sum_{b=1\ldots d}\frac{\|x_i^b - x_j^b\ \|^2}{2l_b^2}) \tag{8}$$

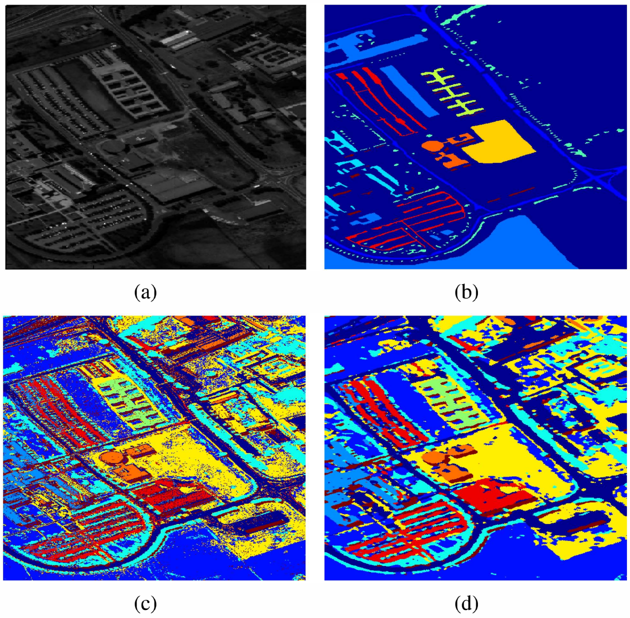
x_i^b indicates the bth band of the ith input point. The ARD has been proved to be an effective kernel successfully removing irrelevant information [12]. It provides a parametrization scheme for automatic feature reduction especially for the high-dimensional challenge such as HSI with more than one hundred bands.

B. MRF-based Regularization

In the aforementioned pixelwise classification, only the spectral information is considered. However the spectral response can be affected by other spectrum from adjacent pixels. Therefore, it is necessary to regularize the pixelwise classification results with MAP-MRF framework [13].

Algorithm 1 GP-MRF

```
Input: $P_L(x_i|y_i)$: the the likelihood function for pixel $i$ belonging to a class $y_i$;
       $Im$: the label map from GPC;
Output: optimal $y^* \rightarrow$ new label map
initial the minimal global energy $E_{min}$;
compute spectral energy $E_{spectral}$;
find the neighbourhoods $\mathcal{N}$ for each site;
repeat
    compute spatial energy $E_{spatial}$ based on $Im$;
    compute local energy $E(y_i)$ for each site;
    assign the new label $y_i^*$ corresponding to $\min E(y_i)$ to the site $i$ and update label map $Im$;
    compute the global energy $E(y)$ and compare with $E_{min}$;
    if $E(y) \leq E_{min}$ then
        $E_{min} \leftarrow E(y)$
    end if
until $E_{min}$ convergence
```



(a) (b) (c) (d)

Fig. 1. (a) Data of the University of Pavia, (b) ground truth, (c) classification result of GPC (ARD) and (d) classification result of GP-MRF (ARD).

Markov Random Fields are a probabilistic framework that incorporates the spatial information from a set of cliques in images, whose basic principle is that each pixel interacts only with its neighbouring pixels [7]. In other words, a pixel more possibly has the same label as its neighbourhoods. Because of formulating MRF models within Bayesian framework, the optimal solution is the *Maximum a Posteriori* (MAP) and is obtained by maximizing the posterior probability $Pr(\mathbf{y}|\mathbf{x})$:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} Pr(\mathbf{y}|\mathbf{x}) \tag{9}$$

where $\mathbf{x}$ is the observation and $\mathbf{y}$ is the possible labeling.

Based on the *Hammersley-Clifford theorem*, we consider the MAP solution as the minimization of an energy cost function [14]:

$$E = E_{spectral} + E_{spatial} \tag{10}$$

$E_{spectral}$ is the spectral energy defined by the likelihood function as:

$$E_{spectral} = -ln\{P_L(x_i|y_i)\} \tag{11}$$

where x_i is the site of the ith pixel in the label map, y_i is one of the possible label for site x_i, and the likelihood function $P_L(x_i|y_i)$ have been already yielded by GPC (i.e. $P(y_i|\mathbf{x}_i)$), which means the predictive probability of x_i belonging to the class y_i. The second term of Eq. (10) is spatial energy and its standard expression is:

$$E_{spatial} = \sum_{j\in\mathcal{N}} \beta(1-\delta(y_i,y_j)), j \in \mathcal{N} \tag{12}$$

where $\delta(.,.)$ is the Kronecker delta function ($\delta(a,b) = 1$ if $a = b$, else $\delta(a,b) = 0$) and β is a non-negative parameter controlling the weight of spatial energy. $\mathcal{N}$ is the neighbourhood system, which in this paper is 8-connected. y_i is the label of the center pixel x_i and y_j is the label of its jth neighbouring pixel.

We adopt the *Iterative Conditional Modes* (ICM) [15] to solve the optimization problem. We compute the local energy $E(x_i)$ of each pixel belonging to each label. A pixel is assigned to the label with smallest energy and it gets the local optimization. The local energies were summed up as global energy. Based on the updated label map the above procedure will be repeated. The optimization can be achieved until the global energy is convergence. The procedure is detailed in Algorithm 1.

III. EXPERIMENTAL RESULTS

In the experiments, three hyperspectral datasets-INDIAN PINES, UNIVERSITY OF PAVIA, and CENTER OF PAVIA will be used in this paper. These datasets have been widely used as benchmark [2]-[5] in the study of HSI classification. The INDIAN PINES dataset was acquired by the AVIRIS in 1992 and taken over a predominately agricultural region in NW Indiana, USA. The dataset has 145×145 pixels and 200 channels. Seven of the 16 different land-cover classes in the original ground-truth were removed, which can offer only a few training samples (this makes the experimental analysis more significant from the statistical viewpoint) [2]. The CENTER OF PAVIA image remains 102 channels after removing some noisy bands and lies around the center of Pavia with 1096×492 pixels. The ground-truth consists of 9 land-cover classes. The UNIVERSITY OF PAVIA dataset has 103 channels with 610×340 pixels and also 9 land-cover classes.

In the experiments, both the RBF and ARD kernel were adopted in the GP model for comparison purpose and the hyperparameters were optimized by *Conjugate Gradient* method [16] based on the *Laplace* method. In order to simplify the classification and balanced samples problems, the one-against-one strategy was applied. The algorithm [17] was used to estimate the predictive probability of the test samples belonging to each class from the results of one-against-one strategy.

The original image and ground truth of the University of Pavia dataset are shown in Fig.1(a) and Fig.1(b) respectively. The classification results of GPC are shown in Fig.1(c). Many scattered pixels or small patches are labeled as different classes from their adjacent pixels by GPC. These labels are unconvinced as we have discussed in Section I. Fig. 1(d) shows the improved classification results by MRF based on the results of GPC. The label image is refined by MRF. In this experiment, the ARD kernel was used in GP model. We used the same size of training and test samples as in [18].

Table I shows the individual class accuracy of SVM, RF, GP (RBF), GP (ARD) and GP-MRF (ARD) from the

TABLE I. INDIVIDUAL CLASS PERCENTAGE ACCURACIES OF THE UNIVERSITY OF PAVIA DATASET WITH DIFFERENT CLASSIFIERS.

Class	SVM	RF	GP_{RBF}	GP_{ARD}	GP-MRF
C1 Asphalt	85.4	84.7	88.5	91.1	**99.2**
C2 Meadows	65.9	90.9	94.5	94.3	**99.2**
C3 Gravel	68.8	86.9	90.0	89.2	**97.7**
C4 Trees	97.0	95.1	97.4	**97.4**	97.0
C5 Metal Sheets	99.4	99.6	**100**	**100**	**100**
C6 Bare Soil	93.7	65.8	87.5	89.0	**98.6**
C7 Bitumen	90.5	91.3	93.1	95.2	**98.5**
C8 Bricks	92.5	70.9	78.9	82.4	**92.1**
C9 Shadow	97.5	**100**	99.8	**100**	99.5

TABLE II. OA AND AA IN PERCENTAGE OF GP (RBF), GP (ARD) AND GP-MRF (ARD) FOR DIFFERENT DATASETS

Algorithm	INDIAN		UNIVERSITY		CENTER	
	OA	AA	OA	AA	OA	AA
GP (RBF)	84.50	89.39	90.58	91.46	98.33	96.53
GP (ARD)	87.26	91.41	92.25	93.15	**98.41**	96.60
GP-MRF (ARD)	**95.60**	**97.42**	**98.3**	**97.9**	97.48	**99.13**

University of Pavia dataset. In order to objectively compare the performances between different classifiers, we used the same size of training and test samples as [18] and quoted the experimental results of SVM (RBF). The results show that the GPC performs competitively or even better than the state-of-the-art methods SVM and RF in terms of accuracy. The comparison between the GPC (RBF) and GPC (ARD) proves the previous discussion in Section II: the ARD kernel outperforms RBF kernel for classification of HSI. However, in order to optimize more parameters for ARD kernel, more input dimensions increase the training time rapidly. Finally, the results of GP-MRF (ARD) demonstrate that our proposed approach can significantly increase the classification accuracy of the individual class.

Table II compares the results in terms of overall accuracy (OA) and average accuracy (AA) between GP (RBF), GP (ARD) and GP-MRF (ARD) in three different datasets. The results further prove that, our proposed approach can effectively improve the accuracies of classification for HSI over urban/suburban regions. 200 points for each class from these datasets were randomly selected as training samples and the residual were regarded as test samples.

Finally, Fig. 2 investigates the performances of GP-MRF (ARD) in terms of global classification accuracy with different weight parameter $\beta = [0.5, 1, 2, 3, 4, 5]$ for spatial information in Eq. (12). We draw the conclusion that the OA is not significantly different over the given values. Our method is robust to the choice of β.

IV. CONCLUTION

In this paper we proposed a novel framework GP-MRF, which combines the GPC and MRF to enhance the classification accuracies. The GP-MRF framework integrates the spectral information into spatial information and effectively classifies the HSI over urban/suburban regions without selection or reduction of data dimensionality.

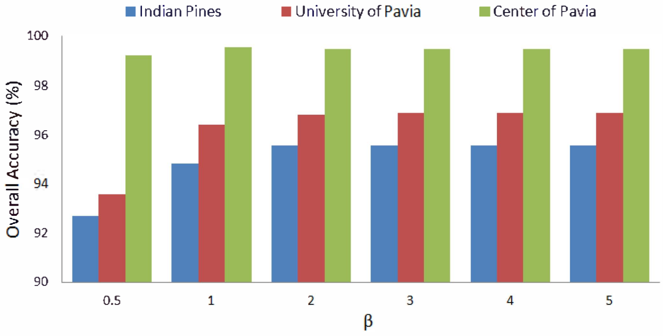


Fig. 2. Overall accuracy in percentage for different values of β for different datasets

We evaluated the performance of GP-MRF in three hyperspectral datasets and the results demonstrated that MRFs utilize the relationship between the adjacent pixels to improve the classification accuracy of HSI on the basis of GPC classification. We used GPC to preliminarily classify original data and obtain label image and predictive probability of each pixel belonging to each class which will be applied in the step of MRF. The experiment shows that our approach yields accurate classification results and is robust for classifying different kinds of HSI.

REFERENCES

[1] G. Hughes, On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory, Vol. 14, No. 1, pp. 55-63, 1968.

[2] F. Melgani and B. Lorenzo, Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing, Vol. 42, No. 8, pp. 1778-1790, 2004.

[3] J. Ham, Y. Chen, M.M. Crawford and J. Ghosh, Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 3, pp. 492-501, 2005

[4] G. Camps-Valls and L. Bruzzone, Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 6 , pp. 1351-1362, 2005.

[5] K. Zhao, S. Popescu and X. Zhang, Bayesian learning with Gaussian processes for supervised classification of hyperspectral data. Photogrammetric Engineering and Remote Sensing, Vol. 74, No.10, pp. 1223-1234, 2008.

[6] S. Kumar, Models for learning spatial interactions in natural images for context-based classification. Ph.D. thesis, Carnegie Mellon University, 2005.

[7] S.Z. Li, Markov random field modeling in image analysis. Vol. 26. Springer, 2009.

[8] A.H.S. Solberg, T. Torfinn and A.K. Jain, A Markov random field model for classification of multisource satellite imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 34, No. 1, pp. 100-113, 1996.

[9] C.E. Rasmussen, C.K.I. Williams, Gaussian processes for machine learning. The MIT Press 2006.

[10] C.K. Williams and B. David, Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, pp. 1342-1351, 1998.

[11] T.P. Minka, A family of algorithms for approximate Bayesian inference. Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[12] C.E. Rasmussen, C.K.I. Williams, Gaussian processes for regression. The MIT Press 1996.

[13] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, No. 6, pp. 721-741, 1994.

[14] M.Y. Yang and W. Förstner, A hierarchical conditional random field model for labeling and classifying images of man-made scenes. ICCV Workshop on Computer Vision for Remote Sensing of the Environment, pp. 196-203, 2011.

[15] S.J. Prince, Computer vision: models, learning, and inference. Cambridge University Press, 2012.

[16] J. Nocedal and S.J. Wright. Numerical Operation. pp. 101-134, Springer, 2006.

[17] T.F. Wu, C.J. Lin and R.C. Weng, Probability estimates for multi-class classification by pairwise coupling. The Journal of Machine Learning Research, Vol. 5, pp. 975-1005, 2004.

[18] R. Roscher, B. Waske and W. Förstner. Incremental import vector machines for classifying hyperspectral data. Geoscience and Remote Sensing, IEEE Transactions on Vol. 50, No. 9, pp. 3463-3473, 2012.