Информационные процессы, Том 2, № 1, 2002, стр. 12–21
© 2002 Teterukovskiy, Yu.

Contextual Reclassification of Multispectral Images: A Markov Random Field Approach

Alex Teterukovskiy and Jun Yu

Department of Forest Resource Management, The Swedish University of Agricultural Sciences, S-901 83 Umeå, Sweden

Received April 10, 2002

Abstract—This work presents methods for multispectral image classification using the contextual classifiers based on Markov Random Field (MRF) models. Performance of some conventional classification methods is evaluated, through a Monte Carlo study, with or without using the contextual reclassification. Spatial autocorrelation is present in the computer-generated data on a true scene. The total misclassification rates for varying strengths of autocorrelation and for different methods are compared. The results indicate that the combination of the spectral-contextual classifiers can improve to a great extent the accuracy of conventional non-contextual classification methods. It is also shown how the most complicated cases can be handled by the Gibbs sampler.

Key words: Monte Carlo study, contextual classification, Markov random field, ICM, Gibbs sampler, spatial autocorrelation, multi-spectral imagery, remote sensing.

1. INTRODUCTION

The objective of the supervised classification in remote sensing is to partition the noisy image into classes (e.g. forest and non-forest, forest types with respect to species composition and age), with the parameters in the model estimated from the training samples. Usually, the classification is based on the spectral signature of the classes. One should be aware, however, that the multispectral image data acquired by remote sensing devices are very complex entities. They have not only spectral attributes (with correlated spectral bands) but also spatial attributes. Proper utilization of this spatial contextual information, in addition to spectral information, usually will significantly improve the classification performance compared to the conventional noncontextual rules such as the linear and quadratic discriminant analysis (LDA and QDA) and the k-nearest neighbour (k-NN) classifier.

In noncontextual classification rules, the spatial dependence is not accounted for. These approaches effectively assume the spectral intensities in neighbouring pixels to be independent, therefore neglecting the important information from the neighbourhood. Such approaches might be reasonable if the pixel-sizes are large, or when the densities of the spectral intensities are well separated for different classes. In forestry applications, for example, the densities of the spectral intensities are seldom well separated.

In order to take the information from the neighbouring pixels into account, many efforts have been made on the spatial contextual classification methods during the past two decades. In particular, such methods have been introduced to cope with segmentations and classifications of remotely sensed data (see [3, 4, 10, 12] and therein). These approaches tend to be computationally intensive, due to either the estimation of autocorrelation parameters and transition probabilities in pixelwise classification (such as Haslett's method studied in [4]) or a number of iterations in simultaneous classification of all pixels (such as the ICM algorithm studied in [3]). For the pixelwise contextual classification methods, there is a high risk that the spatial correlation parameters can not be properly estimated, see for example the simulation study

in [14]. One more shortcoming of the pixelwise contextual methods is the requirement on the training data that is to be of the form of (homogeneous) crossings, which is not easily accessible in practice.

Despite the computational effort they require, contextual classifiers that rely on macro models for the behaviour of the classes, such as Markov random fields, are of great interest as a smoothing process over an initial classification image. Previous studies [2, 3, 8] have shown that the iterated conditional modes (ICM) method has a promising performance on achieving high accuracy classifications, with less cost on computational effort compared to other Bayesian based contextual classifiers.

As mentioned in [8], it is difficult to assess the properties of the various classification methods by direct theoretical means. In this paper we decided to make a simulation study for evaluation of the ability of contextual classifiers (the ICM and the Gibbs sampler) to improve the initial classifications based on QDA and k-NN. We compare various combinations of spectral-contextual classifiers on a scene (ground truth) where the two-dimensional spectral intensities, given classes, have varying degrees of spatial dependency, from independence to strong autocorrelation. Moreover, the data is generated in such a way that the spectral intensities are highly overlapped. Therefore, the training sets will be highly overlapping in the representation space. We shall concentrate our interest on the misclassification rates they provide.

This paper is organized as follows. In Section 2 we introduce the classifiers. The simulation experiment for creating the simulated data sets, including the autocorrelation model and the true scene, is described in depth in Section 3. The obtained classification results are described in Section 4, while Section 5 contains the discussion and conclusions.

2. METHODS

We assume that a given pixel from the scene belongs to one of a fixed number of classes, say, $1, \ldots, K$. The proportion of pixels belonging to class c in the population under study is denoted by π_c, which is usually unknown. Each pixel gives rise to certain measurements (e.g., spectral signatures in remote sensing images, or their various transformations), which form the feature vector $\mathbf{X}$. Our task is to allocate each pixel into one of the classes in $\{1, \ldots, K\}$, on the basis of the observed value $\mathbf{X} = \mathbf{x}$ and/or prior information about the classes.

2.1. Conventional classification methods

In this study, two conventional classification methods are used: QDA and k-NN, where the former is known as a parametric classification rule and the latter as a nonparametric rule. For details we refer to, among many others, Ripley's book [9]. Note that these methods assume independency between pixels.

Parametric model

In parametric models, the feature vectors from class c are assumed to be distributed according to the density $p_c(\mathbf{x})$. Then the posterior distribution of the classes after observing $\mathbf{x}$ is

$$p(c|\mathbf{x}) \propto \pi_c p_c(\mathbf{x}).$$

If we assume the probability model in which the observations for class c are (d-dimensional) multivariate normal with mean $\boldsymbol{\mu}_c$ and covariance matrix Σ_c, the Bayes rule is to allocate a future observation $\mathbf{x}$ to the class that minimises

$$Q_c = -2 \log p(c|\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_c)^T \Sigma_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) + \log |\Sigma_c| - 2 \log \pi_c.$$

This method is known as *quadratic discriminant analysis* due to the quadratic form of $\mathbf{x}$ it involves.

When the classes have a common covariance matrix Σ, Q_c becomes a linear function of $\mathbf{x}$ plus a quadratic term which does not depend on the class. So minimizing Q_c is equivalent to maximizing the linear terms

$$L_c = 2\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_c - \boldsymbol{\mu}_c^T\Sigma^{-1}\boldsymbol{\mu}_c + 2\log\pi_c,$$

which leads to the *linear discriminant analysis*.

In practice, we replace $\boldsymbol{\mu}_c$, Σ_c or Σ, and π_c with their estimates (e.g., maximum likelihood, method of moment, etc.) by using a training sample.

Nonparametric model

There is a number of nonparametric classifiers based on nonparametric estimates of the class densities or of the log posterior, such as kernel methods, orthogonal expansion, projection pursuit, and others. Here we use the simple adaptive kernel method which gives a classifier known as the *k-nearest neighbour rule*. It is based on finding the k nearest (in the Euclidean distance) pixels from the training samples, and taking a majority vote amongst the classes of these k samples, or equivalently, estimating the posterior distributions $p(c|\mathbf{x})$ by the proportions of the classes amongst the k samples.

The k-NN classifier uses the whole training sample set of any class to choose the neighbourhood, which differs from using the k-nearest neighbour density estimate for each class. Ties in distances can occur with finite-precision data. One solution to this problem is to include in the vote all distances equal to the kth largest.

2.2. Contextual Bayesian classification

By contextual classification it is understood that the class of a given pixel is based both on the spectral value of this pixel and the spectral values and/or classes of its spatial neighbours. In most classification tasks where some sort of structure can be observed in the image, this additional information proves to be very important for the overall quality of classification. In order to include information from the neighbours it is necessary to specify a joint model for the behaviour of all classes. One common choice for such model is a Markov random field (MRF). In this model one supposes that the class of pixel $\mathbf{s}$, given the classes in the rest of the image, depends only on the classes in the neighbours $\partial\mathbf{s}$ of $\mathbf{s}$. Naturally, by extending the neighbourhood to the whole image, one can consider *any* spatial random model as a Markov random field. If the neighbourhood is taken small enough, and the dependence between neighbours is chosen wisely, the inference on the image using MRF model becomes possible. Classification results obtained by these methods effectively outperform the results of non-contextual classifiers.

We have chosen the following form of dependence between neighbouring classes [1]. The second-order neighbourhood $\partial\mathbf{s}$ is considered, and the probability

$$P(C(\mathbf{s}) = c) \propto \exp\left(\beta \times \sum_{\mathbf{t}\in\partial\mathbf{s}} I(C(\mathbf{t}) = C(\mathbf{s}))\right), \tag{1}$$

where $I(\cdot)$ is the indicator function. Using this representation for the prior distribution of the classes we can write the posterior distribution as

$$p\left(c|\mathbf{x}\right) \propto \Phi_c(\mathbf{x}) \times \exp\left(\beta \times \sum_{\mathbf{s}} \sum_{\mathbf{t}\in\partial\mathbf{s}} I(C(\mathbf{t}) = C(\mathbf{s}))\right).$$

Here $\Phi_c(\mathbf{x})$ denotes the Gaussian distribution with mean and covariance matrix depending on the classes. The parameter β indicates the strength of attraction within classes.

Maximization of the posterior distribution gives us the maximum a posteriori estimate (MAPE). Unfortunately, direct maximization of the distribution $p(c|\mathbf{x})$ requires unrealistic computational effort. Therefore, several algorithms for approximation of MAPE were suggested [13]. Two approaches are considered in this paper: one quickly giving the *local* maximum of the posterior distribution, and another - approaching the *global* maximum but on the expense of CPU time.

Iterated conditional modes

The ICM method was described by Besag [1], and is, in fact, a statistical reformulation of the coordinate-wise maximal descent method well-known in optimization. It is often used in image analysis applications since it is fast and easy to implement. Instead of maximizing $p(c|\mathbf{x})$ the maximization of the marginal distributions $p\left(C(\mathbf{s})|\mathbf{x}\right)$ is considered. For each pixel $\mathbf{s}$ such marginal distribution is concentrated in the number of atoms equal to the number of classes. Choosing the mode of this distribution, and then repeating the procedure for all pixels (possibly, several times) will soon produce a configuration which does not have neighbours with higher posterior probability. Such configurations are by definition the local maxima of the posterior distribution.

Since this method searches for the local maximum, it strongly depends on the initial configuration. This is a drawback on one hand, but can also be seen as a guarantee that the classification will not deviate strongly from the true one if the prior distribution is chosen somewhat incorrectly. In cases where we want to obtain the global maxima of the posterior distribution, the ICM is inappropriate. A simple modification of this algorithm provides (at least, theoretically) means to obtain a global maximum.

The Gibbs sampler and simulated annealing

Instead of choosing the mode of the marginal distribution, Geman and Geman [5] suggested sampling from it. They called their algorithm the Gibbs sampler since it has an obvious connection with the Gibbs fields in statistical physics. They have proven that provided all pixels are visited frequently enough, the stationary distribution of the Gibbs sampler is the posterior distribution $p(c|\mathbf{x})$. For details about the Gibbs sampler we refer to the original paper and a book by Winkler [13].

We shall use the modification of the Gibbs sampler called *simulated annealing*. The idea of this method is to emphasise the modes of the posterior distribution by modifying the conditional distributions which are used in the Gibbs sampler in the following way. Let us write out the targeted posterior distribution in the form

$$p(c|\mathbf{x}) = \frac{1}{Z}\exp(-H(c,\mathbf{x})),$$

where Z is the normalizing constant. Then the distribution

$$p^T(c|\mathbf{x}) = \frac{1}{Z^T}\exp(-\frac{H(c,\mathbf{x})}{T}), \quad T>0$$

will for small T have the same maxima as $p(c|\mathbf{x})$ but easier detectable. This observation is the basis of the simulated annealing. The parameter T is called the *temperature*. If, in the course of running the Gibbs sampler, the temperature is lowered according to a certain *cooling schedule*, then the limit distribution of the sampler will be concentrated on the set of maximizers of the posterior distribution. When we in the section devoted to results refer to the Gibbs sampler, it is always understood that the sampler is coupled with simulated annealing with a linear cooling schedule.

It is necessary to mention that the Gibbs sampler converges much slower then the ICM algorithm - typically 400 $\sim$ 1000 sweeps versus 5 or 6. Therefore, for classification problems it is useful to consider Gibbs sampler if one is certain about the adequacy of the prior distribution.

3. AUTOCORRELATION MODEL AND THE TRUE SCENE

In this section we describe how the data that we used to illustrate the classification procedures introduced in Section 2, were simulated. The true scene and the model for the spectral intensity are similar to those used in [8], where the performance of different contextual classification methods was evaluated.

The stochastic model for the d-dimensional spectral intensity, $\{X(\mathbf{s}), \mathbf{s}\in\mathcal{R}^2\}$, is assumed to be stationary and isotropic, with the following structure:

$$X(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{2}$$

where $\varepsilon(\mathbf{s})$ is the autocorrelated noise process which accounts for the class independent variations, and $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$ are the "true" intensities, conditionally independent given their classes, and also independent of the noise process. The distribution of $Y(\mathbf{s})$ given its class at pixel $\mathbf{s}$, $C(\mathbf{s}) = c$, is multivariate normal with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $(1-\theta)\lambda_c\Gamma$, where θ and λ_c's are proportion parameters and Γ is a fixed $d \times d$ positive definite matrix. Thus, the covariance matrices of Y are allowed to be different for different classes. The assumption of the distribution for the noise vectors $\varepsilon(\mathbf{s})$ is as follows: $\varepsilon(\mathbf{s}) \sim N(\mathbf{0}, \theta\Gamma), Cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{t})) = \rho^{|\mathbf{s}-\mathbf{t}|}\theta\Gamma$, where $|\mathbf{s}-\mathbf{t}|$ is the Euclidean distance between pixels $\mathbf{s}$ and $\mathbf{t}$, and ρ is the spatial autocorrelation between the noise vectors.

Hence, model (2) implies that the spectral intensity X is multivariate normal with

$$\begin{aligned} E(X(\mathbf{s}) \mid C(\mathbf{s}) = c) &= \boldsymbol{\mu}_c \\ Var(X(\mathbf{s}) \mid C(\mathbf{s}) = c) &= \Sigma_c = [(1-\theta)\lambda_c + \theta]\Gamma \\ Cov(X(\mathbf{s}), X(\mathbf{t}) \mid C(\mathbf{s}), C(\mathbf{t})) &= Cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{t})) = \rho^{|\mathbf{s}-\mathbf{t}|}\theta\Gamma. \end{aligned}$$

Although model (2) with different values of λ_c was used in [8] for generating the data, they actually set $\lambda_c \equiv 1$ when performing the classification. This model was also studied in [4] with specification of $\lambda_c \equiv 1$, i.e., equal covariance matrix for different classes.

In this study, we consider a true scene of size 40×40 pixels, as shown in Figure 1. This picture is, in fact, a small selection from a real scene. The number of classes is $K = 3$.

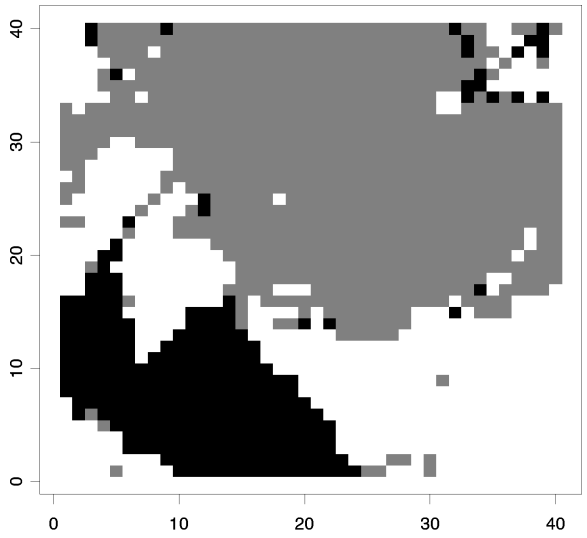

Fig. 1. The true scene. Legends: Class 1 = black; Class 2 = white, Class 3 = grey.

To specify the model (2) for generating our data, we follow mostly the values in Mohn et al. [8]. Consider an image with two correlated bands, that is, $d = 2$. We choose the values of λ_c's as 0.4, 1.4, and 1.0, respectively, which gives different covariance matrices for the three classes. The class expectations and the matrix Γ are chosen as

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 125 \\ 128 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} 130 \\ 135 \end{pmatrix}, \ \boldsymbol{\mu}_3 = \begin{pmatrix} 127 \\ 110 \end{pmatrix}, \ \ \Gamma = \begin{pmatrix} 4 & 8 \\ 8 & 81 \end{pmatrix}.$$

As a measure of the spatial dependency of the noise term we use the nearest neighbour correlation for class 3, $\alpha = \rho\theta$, as specified in the following table.

α	0	0.25	0.72
ρ	0	0.5	0.9
θ	0.5	0.5	0.8

Thus, $\alpha = 0$ implies spatial independency between pixels, which is the case most appropriate for the noncontextual classification methods. The value of the autocorrelation parameter ρ depends on the pixel size. The larger the pixels are, the smaller is the value of ρ. We refer to [6] for an illustration of this relationship.

Figure 2 shows the location of the three distributions in the case $\alpha = 0$. The ellipsoids are centered in the expectations. Many of those pixels with spectral vectors falling in the intersection between classes will be misclassified.

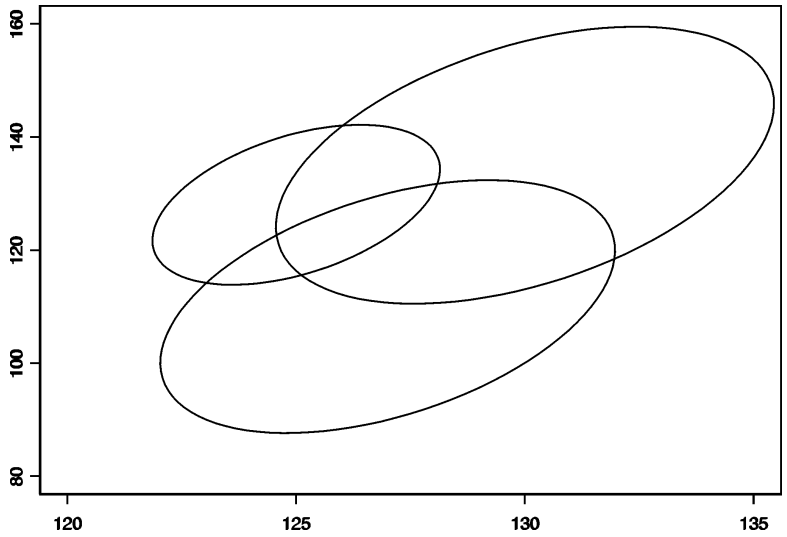

Fig. 2. Class expectations ± 2 standard deviation contours. Case $\alpha = 0$.

4. RESULTS

In this Monte Carlo study, 50 replicates of the true spectral intensity $Y = \{Y_i, i = 1, \ldots, 50\}$ and 50 replicates of the spatially correlated noise $\varepsilon = \{\varepsilon_i, i = 1, \ldots, 50\}$ were generated, for each particular choice of α-value. Recall that both Y_i and ε_i are 2-dimensional with correlated components. Then 50 replicates of the 2-dimensional noisy image X were constructed as follows: $X = \{Y_i + \varepsilon_i, i = 1, \ldots, 50\}$. For training data, we follow the approach of Klein and Press [7], by taking observations from all the pixels in some columns. Specifically, columns 3, 8, 13, 18, 23, 28, 33, and 38 were chosen, which constitute 1/5 of the scene. For each replication, the parameters $\boldsymbol{\mu}_c$ and Σ_c in the model were estimated (using maximum likelihood) and plugged into the classification rules, and all pixels, except those in the training set, were classified. The results of initial classifications by k-NN and QDA are then used as the inputs for the ICM and the Gibbs sampler. The parameter β in (1) is chosen to be fixed at 0.75 over all iterations. The choice of this parameter is somewhat of an expert knowledge. The model for classes that we adopted is in fact similar to Potts model, which is in its turn a generalization of the Ising model in statistical physics. By comparing our true scene with realizations of the Ising field we were able to make a realistic choice of β. In real life applications, where the true scene is unknown, this problem is not so easily solved. The number of iterations is fixed in advance at 8 for the ICM reclassification. The implementation of the simulated annealing with the fast cooling schedule allowed us to stop the Gibbs sampler after only 100 iterations.

The statistical software S-PLUS is used throughout the simulation study in Section 3 and 4. The MASS library [11] is employed for the purpose of driving the conventional classification methods (k-NN and QDA).

4.1. Total misclassification rates

The total misclassification rates are estimated by averaging the relative number of misclassified pixels in each replication. Table 1 shows the mean values of misclassification percentages. The first column indicates the three different structures of spatial dependency ($\alpha = 0, 0.25$, and 0.72, respectively). The last three columns represent the initial classification methods mentioned in Section 2.1, where the prior distribution used in QDA was chosen to be the class proportion coming from the training data, and the number of neighbours in k-NN was chosen to have two different alternatives: $k = 1$ or 3, in order to check the neighbouring effect. In each cell the first and second rows show the misclassification rates of the initial classification and of the ICM reclassification, respectively. The best results at each spatial dependence level are indicated with bold face.

Table 1. *Mean misclassification rates (%)*

Spatial dependency		1-NN	3-NN	QDA
$\alpha = 0$	Original	19.36	16.44	13.18
	ICM	3.86	3.84	**3.80**
$\alpha = 0.25$	Original	18.95	16.20	13.12
	ICM	4.58	**4.56**	**4.56**
$\alpha = 0.72$	Original	20.82	17.87	14.67
	ICM	**6.98**	7.10	7.15

What we can observe and conclude from Table 1 is as follows:

- The ICM reclassification can improve to a great extent the accuracy of k-NN and QDA classifiers. The mean misclassification rate from the initial classification based on these non-contextual classifiers can be reduced up to 80%.
- For the initial classification, QDA worked better than the k-NN classifier, especially in the case when $k = 1$. This is quite natural because our model is multivariate normal which fits the basic requirements in QDA. But the performance of the initial classification seems not to affect our final results. This is explained by the fact that the ICM classifier seeks the local maximum of the posterior distribution, starting from all initial classifications not far from this local maximum, however distinct they may be.
- In general, the mean misclassification rate increases as the spatial dependency between pixels strengthens.

Table 2 complements Table 1 in the following way. It is organized in the same manner but instead of the mean misclassification rates, the standard deviations of those are reported. The lowest standard deviations at each dependence level are indicated with bold face.

Table 2. *Standard deviations of mean misclassification rates (%)*

Spatial dependency		1-NN	3-NN	QDA
$\alpha = 0$	Original	1.38	1.41	1.00
	ICM	0.44	**0.42**	**0.42**
$\alpha = 0.25$	Original	2.18	1.74	1.12
	ICM	0.74	0.68	**0.62**
$\alpha = 0.72$	Original	6.12	5.61	4.76
	ICM	**2.36**	2.59	2.65

It follows from Table 2 that the standard deviation increases with the spatial dependency. It should also be noticed that the lower the misclassification rate is, the lower is the standard deviation.

We have also investigated the ability of the Gibbs sampler to maximize the posterior distribution. Figures 3–5 show the results of the implementation of the Gibbs sampler for reclassification of the images with

strongest autocorrelation (0.72), whose initial classification was obtained by QDA - the best noncontextual method among those that we considered. For illustrative purposes, we display the results that the Gibbs sampler attained in the following three distinct groups of replicates.

– Figure 3: Group 1 (worst) - a collection of three replicates that proved most difficult for QDA. This means that the initial classification by QDA of those replicates had the highest misclassification rate (26.28% in average).

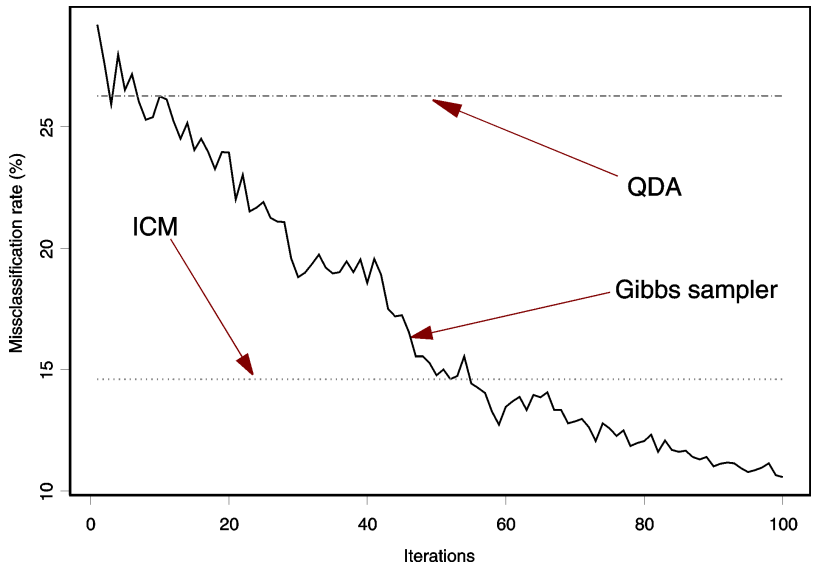

Fig. 3. Misclassification rates for QDA, QDA+ICM and QDA+Gibbs sampler for bad initial classification.

– Figure 4: Group 2 (average) - three replicates for which the initial classification by QDA produced intermediate misclassification rates (14.66% in average).

– Figure 5: Group 3 (best) - three replicates that were classified by QDA with lowest misclassification rates (8.39% in average).

The obtained results are as follows. For images with poor initial classification, the reclassification by the Gibbs sampler is preferable to that by the ICM. This is explained by the fact that the bad initial classification is located far (in the image space) from the global maximum of the posterior distribution, and therefore such algorithm as the ICM that searches for a local maximum can not be expected to perform well. For images with average and good initial classifications, the implementation of the Gibbs sampler leads to unnecessary loss of CPU time. As one can see in Figure 5, for images with the best initial classification, the Gibbs sampler has not attained even the local maximum after a 100 iterations.

5. DISCUSSION AND CONCLUSION

The results of our Monte Carlo study indicate that the contextual classifiers as a post smoothing process after the initial classification can improve the accuracy of the final classification, even when the multispectral remote sensing data is highly overlapped and strongly spatially correlated.

The Gibbs sampler works better than the ICM when the initial classification is poor. In this case the local maximum of the posterior distribution, found by the ICM, is relatively far from the global one, and therefore the Gibbs sampler allows us to improve the classification rate. On the other hand, if the results of the initial

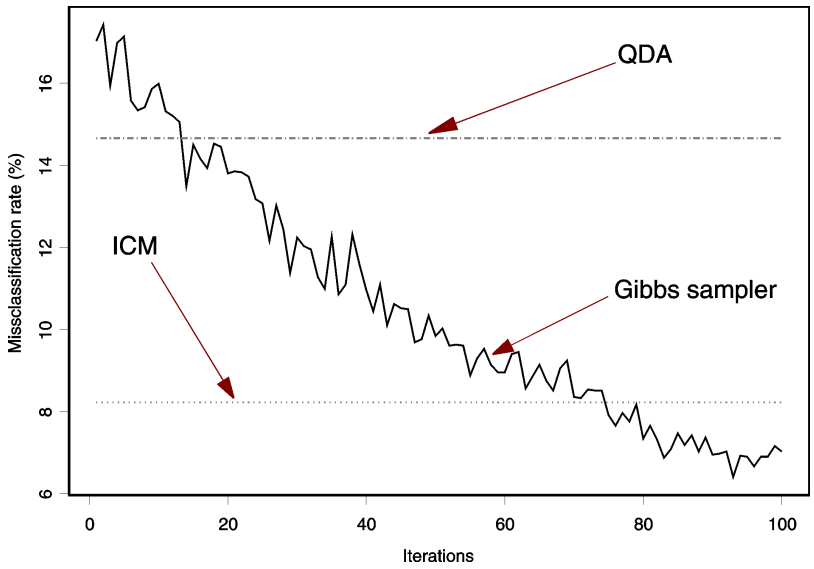

Fig. 4. Misclassification rates for QDA, QDA+ICM and QDA+Gibbs sampler for average initial classification.

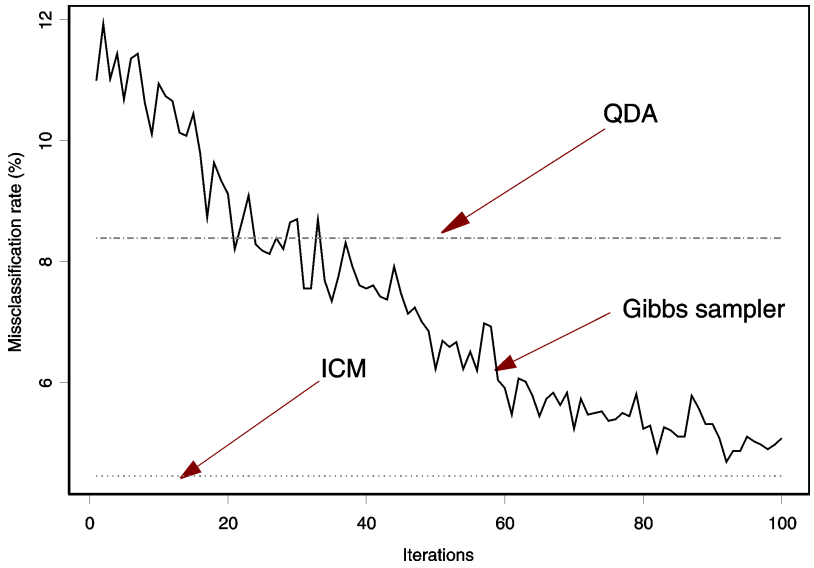

Fig. 5. Misclassification rates for QDA, QDA+ICM and QDA+Gibbs sampler for good initial classification.

classification are satisfactory, the Gibbs sampler is not needed, as there is a large chance that the global maximum of the posterior distribution is the same as the local one, which can be found by the ICM within significantly shorter time.

In this paper, the value of the parameter β is fixed. There is some evidence ([1]) that better results could possibly have been obtained if it had been changed over iterations.

The present study has convinced us that the combination of spectral-contextual classifiers is a creditable approach in the subject of statistical classification of remotely sensed imagery.

We believe that even higher accuracy of reclassification can be achieved for de-noised images. Our initial experiments with the wavelet shrinkage support this conjecture. Further study on this issue will be included in our next report.

Acknowledgement

This work is a part of the Swedish research programme “Remote Sensing for the Environment", RESE (home page http://rese.satellus.se), financed by the Swedish Foundation for Strategic Environmental Research, MISTRA. The authors would like to thank Professor Bo Ranneby and Dr. Magnus Ekstrøm for their valuable comments and discussion.

REFERENCES

1. J. Besag. On the statistical analysis of dirty pictures. *Royal Statistical Society, Ser. B*, 48:259–302, 1986.
2. F. J. Cortijo and N. Perez De La Blanca. A comparative study of multispectral image classifiers: applications to classification problems with high-dimensional data and high-overlapping spectral signatures. simulation of high-dimensional images. Technical Report DECSAI-97-03-04, University of Granada, Spain, 1997.
3. F. J. Cortijo and N. Perez De La Blanca. Improving classical contextual classifications. *Int. J. Remote Sensing*, 19(8):1591–1613, 1998.
4. A-M. Flygare. *Classification of Remotely Sensed Data Utilising the Autocorrelation between Spatio-Temporal Neighbours*. PhD thesis, Department of Mathematical Statistics, UmeaUniversity, 1997.
5. S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.
6. N. L. Hjort, E. Mohn, and G. Storvik. Contextual classification of remotely sensed data, based on an autocorrelation model. *In Sbfi & al.*, 1985.
7. R. Klein and S. J. Press. Contextual bayesian classification of remotely sensed data. *Commun. Statist. - Theory Meth.*, 18(9):3177–3202, 1989.
8. E. Mohn, N. L. Hjort, and G. O. Storvik. A simulation study of some contextual classification methods for remotely sensed data. *IEEE Trans. Geosci. Remote Sensing*, GE-25(6):796–804, November 1987.
9. B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge UK, 1996.
10. K. M. S. Sharma and A. Sarkar. A modified contextual classification technique for remote sensing data. *Photogrammetric Engineering & Remote Sensing*, 64(4):273–280, April 1998.
11. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer, New York, second edition, 1997.
12. T. Watanabe, H. Suzuki, S. Tanba, and R. Yokoyama. Improved contextual classifiers of multispectral image data. *IEICE Trans. Fundamentals*, E77-A(9):1445–1450, September 1994.
13. G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer, Berlin, 1995.
14. J. Yu and M. Ekstrøm. Multispectral image classification using wavelets: a simulation study. Manuscript, 2001.