Unsupervised Classification of Polarimetric SAR Data Using Graph Cut Optimization

M. Jäger, A. Reigber and O. Hellwich

Berlin University of Technology (TUB), Computer Vision and Remote Sensing Group, Franklinstraße 28/29, Sekretariat FR3-1, D-10587 Berlin, Germany.
Tel. +49-30314-73109, Fax. +49-30314-21114, E-mail: jaeger@cs.tu-berlin.de

Abstract—The paper presents a new framework for the classification of polarimetric SAR data. The underlying model introduces cyclic conditional dependencies among the class labels assigned to neighboring observations as a mechanism to regulate the spatial homogeneity of classification results. Classification is posed as an inference problem, and is solved by coherently integrating expectation maximization and graph cut optimization. Results based on real SAR data are presented.

I. INTRODUCTION

The unsupervised classification of synthetic aperture radar (SAR) remote sensing data is a well-studied problem, as is evident from the comprehensive collection of literature on the subject (e.g. see [1] among many others). The classification of polarimetric SAR (PolSAR) data, in particular, has drawn much attention in recent years.

Given a set of \( N_x \) observations \( X = \{ x_1, \ldots, x_{N_x} \} \), the aim of classification is to assign a label \( y_i \in [1, N_C] \), where \( N_C \) is the number of classes, to each observation. The set of labels \( Y = \{ y_1, \ldots, y_{N_x} \} \) then constitutes the classification result, in which label \( y_i \) indicates to which class the observation \( x_i \) has been assigned.

The most established, and perhaps most successful, PolSAR classifiers are based on a statistical model of the observed backscattered signal, and assume that observations \( x_i \) are mutually independent. The classification task then amounts to maximizing the posterior probability of \( Y \), which (noting the absence of prior knowledge in the case of unsupervised classification) is given by

\[
p(Y|X, \Theta) \propto p(X|Y, \Theta) \prod_{i=1}^{N_x} p(y_i | \Theta)
\]

where \( \Theta \) denotes a set of model parameters. The likelihood \( p(x_i|y_i, \Theta) \) typically takes the form of a mixture model, and optimal parameters \( \Theta \) and the corresponding label set \( Y \) can be iteratively inferred using expectation maximization or K-means (see section II-A).

Although the independence implied by the model of (1) is in some sense justified, results are not satisfactory unless the samples \( X \) are carefully pre-processed. The principal cause is the presence of speckle noise. When noise corrupts the observations in two distinct classes, these classes often begin to overlap in the feature space associated with \( X \). If observations are treated independently, such classes can no longer be optimally separated and the classification quality deteriorates. Also, even in the absence of noise, a homogeneous region is not associated with uniform observations. If observations are treated independently, oversegmentation can occur due to naturally occurring in-class variance.

In practice, these problems are overcome by spatially averaging observations to reduce noise and homogenize classification results by implicitly introducing spatial correlations that are not modelled by eqn. (1).

The approach outlined in the following sections does not require spatial homogeneity among samples \( X \). Instead, the model of eqn. (1) is adapted to encourage the spatial homogeneity of labels \( Y \). In doing so, the inevitable decrease in resolution and possible artefacts due to spatial averaging are avoided. Also, the need for considering the spatial context when classifying PolSAR data is made explicit.

The graphical model of figure 1 introduces conditional dependencies among the labels \( y_i \). In this model

\[
p(Y|X, \Theta) = \prod_{i=1}^{N_x} p(y_i|x_i, \Theta)p(y_i|y_{N_i})
\]

where the set \( N_i \) indicates the immediate spatial neighbors of observation \( x_i \), and \( p(y_i|y_{N_i}) \) introduces spatial dependencies among labels.

Maximizing the posterior probability \( p(Y|X, \Theta) \) of (2) is difficult due to the cyclic conditional dependencies among...
labels. For instance, changing the label associated with a single observation directly affects the conditional densities $p(y_i | y_{N_i})$ of neighboring labels, and, by extension, all nodes $y$ in the graph. Typically, the task of inferring a suitable set of labels $Y$ is accomplished by simulated annealing or Gibbs sampling, both of which are extremely computationally intensive. In addition the temperature parameter in simulated annealing is difficult to determine, and Gibbs samplers tend to converge on sub-optimal solutions.

Recently proposed optimization techniques, including belief propagation, normalized cuts and graph-cut optimization, have largely resolved such issues. These developments have motivated the approach outlined in the following sections, in which graph-cut optimization and expectation maximization are combined in a coherent framework for the unsupervised classification of PolSAR data. Section II briefly reviews expectation maximization and graph-cut optimization, while section III presents strategies for effectively combining these techniques. Section IV then presents classification results obtained on real PolSAR data, and V briefly outlines the main contributions.

II. INFERENCE

This section briefly reviews techniques, namely expectation maximization (EM) and graph-cut (GC) optimization, that are useful in the process of inferring the states of latent variables in the graphical model of fig. 1. As explained below, neither inference algorithm is sufficiently powerful to maximize the posterior probability of eqn. (2). Section III describes how the two techniques can be combined to make this possible.

A. Expectation Maximization

The EM algorithm iteratively maximizes the posterior probability of (1), which does not include conditional dependencies among labels. Given an initial set of parameters $\Theta^{(0)}$, the EM iteration yields successively improved estimates $\Theta^{(1)}, \Theta^{(2)}, \ldots$. An optimal label set $Y$ is then easily derived from a given $\Theta^{(t)}$.

In the context of classification, the likelihood $p(x_i | y_i, \Theta)$ of eqn. (1) takes the form of a mixture model,

$$p(x_i | y_i, \Theta) = \sum_{j=1}^{N_C} p(x_i | y_i, \theta_j) \delta(y_i, j)$$

(3)

where each component density $p(x_i | y_i, \theta_j)$ represents the distribution of samples within class $j$, $\Theta = \{\theta_1, \ldots, \theta_{N_C}\}$ is a set of parameters defining these densities, and $\delta$ denotes the Kronecker delta.

After processing, observations in a PolSAR dataset can be represented as scattering vectors of complex co- and cross-polar responses $k = [S_{HH} + S_{VV}, S_{HV}]^T$. Assuming that the observed backscattering has complex Gaussian statistics, the associated covariance matrices $C = \langle k k^\dagger \rangle$, where $\langle \ldots \rangle$ denotes averaging, are known to follow the complex Wishart distribution. In this case, the distribution of samples in class $j$ takes the form

$$p(x_i = C | y_i, \theta_j = \Sigma) = \frac{L^{q/2} |C|^{-q} \exp \left( -L \text{Tr}(\Sigma^{-1} C) \right)}{\Sigma^{L/2} \pi^{(q-1)/2} \prod_{k=1}^L \Gamma(L - k + 1)}$$

where the class parameter $\theta_j$ is the expected covariance $\Sigma_j$ in class $j$, $L$ is the number of looks, $q = 3$, $\text{Tr}(\ldots)$ is the trace of the operand, and $\Gamma$ denotes the gamma function.

Given a set of model parameters $\Theta^{(t)}$, the refinement process begins with the so-called $E$ step, in which the posterior density over labels $Y$ is computed.

$$p(y_i | x_i, \Theta^{(t)}) = \frac{p(x_i | y_i, \Theta^{(t)})}{\sum_{k=1}^{N_C} p(x_i | y_i = k, \Theta^{(t)})}$$

(4)

Intuitively, $p(y_i = j | x_i, \Theta^{(t)}) \in [0, 1]$ represents the degree to which sample $x_i$ belongs to class $j$. These degrees of membership are then, in the so-called $M$ step, used to obtain an updated set of model parameters $\Theta^{(t+1)}$.

$$\theta_j \leftarrow \frac{1}{Z} \sum_{i=1}^{N_S} p(y_i = j | x_i, \Theta^{(t)}) C_i$$

(5)

where $Z = \sum_{i=1}^{N_S} p(y_i = j | x_i, \Theta^{(t)})$, and $C_i$ denotes the covariance matrix associated with observation $x_i$. An optimal label set $Y$ is then obtained by assigning each sample $x_i$ to the class $j$ that maximizes $p(y_i = j | x_i, \Theta^{(t+1)})$.

B. Graph Cut Optimization

Recently proposed optimization techniques [2] using graph cuts produce a label set $Y$ that comes close to minimizing an energy of the form

$$E(Y) = \sum_{i=1}^{N_S} D(y_i) + \sum_{(i,j) \in \mathcal{N}} V(y_i, y_j)$$

(6)

where $\{i, j\} \in \mathcal{N}$ indicates that observations $x_i$ and $x_j$ are immediate neighbors and, equivalently, an edge between nodes $y_i$ and $y_j$ in fig. 1. In (6), the data term $D(y_i = \alpha)$ represents the cost of assigning label $y_i = \alpha$ to observation $x_i$, and encourages a labelling that is consistent with the underlying observations. The smoothness term $V(y_i = \alpha, y_j = \beta)$ penalizes spatial inhomogeneity in the label set $Y$ and introduces cyclical conditional dependencies. In particular, note that defining

$$D(y_i) = - \log(p(y_i | x_i, \Theta))$$

$$V(y_i, y_j) = - \log(p(y_i | y_j))$$

(7)

makes the minimization of eqn. (6) very similar to the maximization of $p(Y | X, \Theta)$ in model (2).

The exact minimization of eqn. (6) wrt. the label set $Y$ is NP-hard if more than two classes are involved. The widely acclaimed article of Boykov et al. [2], however, introduces algorithms of polynomial complexity that minimize the energy of eqn. (6), under certain constraints, to within a bound of the global minimum.

A basic constraint is that set of labels and the associated energies $D(y_i)$ and $V(y_i, y_j)$ must be static and known a priori. This is not the case in the model of fig. 1, where model parameters $\Theta$ are latent and have to be inferred: any useful definition of, at least, the data term $D(y_i)$ must be in terms of class parameters $\theta$.

GC optimization, according to [2], proceeds by constructing a weighted graph with nodes representing observations and
two so-called terminal nodes representing labels. Although
the details of graph construction are beyond the scope of this
review, the weights on graph edges are energies contributed by
different terms in (6). The graph is then cut, by removing edges
with a minimum total weight, in a way that uniquely associates
each observation with one of the terminals. The labeling that
emerges from the cut is then optimal with respect to a subset
of all labels (or classes) in the classification problem.

[2] introduces two styles of graph construction:
\( \alpha \beta \)-swap and \( \alpha \)-extension. Both approaches require that
\( V(y_i = \alpha, y_j = \beta) \geq 0 \) and \( V(y_i = \alpha, y_j = \alpha) = 0 \). The
\( \alpha \beta \)-swap considers all pairs of classes \( \{\alpha, \beta\} \)
in turn, and exchanges labels \( \alpha \leftrightarrow \beta \) in a way that is optimal wrt. (6).
\( \alpha \)-extension considers each class \( \alpha \) in turn, and optimally
enlarges the set of observations assigned label \( \alpha \). \( \alpha \)-extension
requires the smoothness term \( V \) to be a metric, such that
\( V(\alpha, \beta) + V(\beta, \gamma) \geq V(\alpha, \gamma) \). If this condition is met, \( \alpha \)-extension
is to be preferred, since convergence is faster and the minimization
yields an energy with a known upper bound in relation to the global minimum of (6).

III. INTEGRATING EM AND GC OPTIMIZATION

This section describes how the EM and GC optimization
techniques can be combined to allow inference in the model of
fig. 1, where spatial dependencies among labels are considered
and parameters \( \Theta \) are latent.

The overall classification process resembles the EM itera-
tion, in which an initial set of parameters \( \Theta^{(0)} \) is successively
refined until the label set \( \tilde{Y} \) converges to a stable configuration.
To obtain an improved estimate \( \Theta^{(t+1)} \) from \( \Theta^{(t)} \), the E and
M steps of eqns. (4) and (5) are modified to accommodate GC
optimization:

1) The E step no longer computes the posterior densities
over labels \( p(y_i|x_i, \Theta^{(t)}) \), but determines suitable data
and smoothness terms \( D(y_i) \) and \( V(y_i,y_j) \). The data
term is, in the spirit of eqn. (7), consistent with the
maximization of (2).

\[
D(y_i) = -\log \left( p(y_i|x_i, \Theta^{(t)}) \right) \tag{8}
\]
where \( p(y_i|x_i, \Theta^{(t)}) \) is defined in eqn. (4).

2) GC optimization, based on the current \( D(y_i) \)
and \( V(y_i,y_j) \), then yields a label set \( Y^{(t)}_{GC} = \{y_{GC,1}^{(t)}, \ldots, y_{GC,N_y}^{(t)}\} \).

3) Finally, the M step is modified to take \( Y^{(t)}_{GC} \) into account
when a new parameter set \( \Theta^{(t+1)} \) is determined.

In the remainder of this section, III-A describes the simplest
realization of the modified EM iteration outlined above. III-B
and III-C then highlight shortcomings of the basic approach
and introduce refinements.

A. The Pots Model

The so-called Pots model defines a smoothness term that is
user determined and uniform across all classes:

\[
V(\alpha, \beta) = \left\{ \begin{array}{ll}
0 & \alpha = \beta \\
\lambda & \alpha \neq \beta
\end{array} \right. \tag{9}
\]

V is, in this case, a metric and GC optimization can use \( \alpha \)-extension to determine labels \( Y^{(t)}_{GC} \). The M step, in its original
form (5), uses the posterior density over labels \( p(y_i|x_i, \Theta^{(t)}) \).
Since GC optimization yields a concrete labelling, not poste-
rior densities, the M step must use

\[
p^* \left( y_i = j | x_i, \Theta^{(t)} \right) = \delta(\alpha, y_i^{(t)}_{GC,i}) \tag{10}
\]
Note that when \( \lambda = 0 \), indicating no homogenization of labels,
this classifier is identical to the well known \textit{K-means} iteration
used in, for example, [1].

B. Discontinuity Preserving Smoothness

A drawback of the Pots model is that large \( \lambda \), while
promoting noise reduction, can also deform segment outlines
(e.g. by rounding corners) and may even eliminate small
segments entirely.

This issue can, to some extent, be resolved by introducing
adaptive, class dependent energies \( V \). As noted in section
I, the spatial context becomes essential when speckle noise or
insignificant inhomogeneities cause classes to overlap in
feature space. If classes are well separated, regularization in
the form of \( V \) is not necessary and can be dropped in the
interest of avoiding artefacts near segment boundaries.

Having experimented with several adaptive definitions of
\( V \), best results were obtained by quantifying the overlap
(4) \( \alpha \) and \( \beta \) in terms of the posterior densities
\( p(y_i|x_i, \Theta^{(t)}) \) of (4):

\[
(\alpha, \beta) = \sum_{i=1}^{N_y} \min_{c \in \{\alpha, \beta\}} \left( p(y_i = c|x_i, \Theta^{(t)}) \right) \tag{11}
\]
where \( (\alpha, \beta) \in [0, 1] \), and \( (\alpha, \beta) = 0 \) indicates disjoint
classes and \( (\alpha, \beta) = 1 \) indicates complete overlap. The
corresponding definition of \( V \) is given by

\[
V(\alpha, \beta) = \left\{ \begin{array}{ll}
0 & \alpha = \beta \\
\lambda & \alpha \neq \beta
\end{array} \right. \tag{12}
\]
where \( \lambda \) is a user defined parameter that specifies the overall
homogenization, as in (9). Per definition, (12) induces strong
regularization when classes compete for observations and
spatial relations are essential (\( O \to 1, V \to \lambda \)). Boundaries

Fig. 2. The dataset used in the evaluation of the classifier, acquired by the
E-SAR sensor of the DLR over Oberpfaffenhofen, Germany at L-Band.
between clearly disjoint classes, however, are not affected ($O \rightarrow 0, V \rightarrow 0$). $V$ of (12) is not a metric, and the $\alpha/\beta$-swap construction is used for GC optimization.

C. Accumulated Posterior Label Densities

In the original EM algorithm, the posterior $p(y_i|x_i, \Theta^{(t)})$, which is graduated, helps to iteratively resolve uncertainties in the assignment of labels to observations. In contrast, the modified $M$ step of (10) features an essentially binary posterior and does not provide a measure of uncertainty.

A more meaningful posterior can be recovered by accumulating evidence from GC label sets $Y_{GC}^{(t)}$ over several iterations, while using a different value of $\lambda$ in each iteration. Accumulating the evidence in label sets $Y_{GC}^{(t)}$ defines a density

$$p^*(y_i = \alpha|x_i, \Theta^{(t)}) = \frac{1}{t} \sum_{\tau=1}^{t} \delta(y_{GC,i}^{(\tau)}, \alpha)$$

Variability in $\lambda$ is induced by sampling, in each iteration, $\lambda$ from the gamma distribution, such that $\lambda \sim \text{Gamma}(p, \lambda_0)$. $\lambda_0$ indicates the overall degree of homogenization, and precision $p$ determines the variability between iterations.

As in the case of (10), (13) is used to estimate refined parameters $\Theta^{(t+1)}$ in the $M$ step and also to derive a set $Y$ of labels that maximize $p^*(y_i|x_i, \Theta)$. Introducing the accumulated posterior makes the derived label set $Y$ robust with respect to $\lambda$ by removing the need to fix a specific value, and reintroduces the notion of uncertainty in label assignments. More subtly, accumulation effectively decouples EM and GC optimization by allowing EM to converge optimally when $\lambda$ is small, while encouraging homogeneity when it is not.

IV. EXPERIMENTAL RESULTS

Fig. 3 presents classification results obtained using the proposed approach. An initial label set and associated parameters $\Theta^{(0)}$ are obtained from the Freeman-Durden decomposition as described in [1].

As expected, results obtained without considering spatial dependencies among labels are rather noisy due to speckle noise. Introducing such dependencies drastically improves classification quality.

The approaches of section III all produce label sets in good correspondence with the structures evident in the dataset of fig. 2. The refinements suggested in III-B and III-C appear to improve the classification quality, however a quantitative evaluation using ground truth needs to be performed in future.

Importantly, the near-linear time complexity of GC optimization makes the proposed approach only marginally slower than pure $K$-means or EM iteration.

V. CONCLUSION

The paper presents a framework for the unsupervised classification of polarimetric SAR data that combines expectation maximization and graph cut optimization. The underlying model explicitly induces spatial dependencies among class labels, which are shown to result in a significant increase in classification quality.

REFERENCES