Clustering by deterministic annealing and Wishart based distance measures for fully-polarimetric SAR-data

Ronny Hänsch*, Marc Jäger, Olaf Hellwich

*Berlin University of Technology (TUB), Department of Electrical Engineering and Computer Science
Computer Vision and Remote Sensing
Sekr. FR 3-1, Franklinstr. 28/29, D-10587 Berlin
e-mail: rhaensch@fpk.tu-berlin.de
phone: +49 30/314-73107 fax: +49 30/314-21104

Abstract

The basic assumption of clustering is that the observed data embodies a specific structure. This structure is represented by groups of data points, which are relativly isolated within the feature space. The goal of clustering is to find and represent this dissimilar groups of similar elements or, in other words, to subdivide the data space into a number of partitions.
In this paper a spectral clustering based on deterministic annealing is used to obtain image segments with similar spectral properties.
The results are compared with other established clustering methods for polarimetric SAR data and advantages of this approach are investigated.

1 Introduction

Fully-polarimetric SAR messures the amplitude and phase of the backscattered signal in four different combinations of transmit and receive polarisation.
Unfortunately, the speckle effect, due to the coherent nature of the SAR sensor, appears in polarimetric SAR data as well as in common SAR. This strong multiplicative noise often hinders a robust and meaningful segmentation.

In the past different distance measures based on the Wishart distribution were developed, e.g. in [1]. These were used for clustering polarimetric SAR data and deliver more or less stable image segmentations.
However, most algorithms perform hard clustering, that means that one data point belongs to one and only one cluster. Obviously, this might be right for some, but will be wrong for other data points. This should have a negative effect on the clustering result, since forced-choice decisions are hard to undo. Because of the iterative calculations of most algorithms those small errors at the beginning can lead to large errors at the end. Relaxations of the assignment of data points to clusters can be used to defer the clustering decisions, until more evidence is available. This provides a more proper modelling of the uncertainty especially at the beginning of clustering.
Furthermore, most of these methods strongly depend on a proper initialisation e.g. with the H/α-classifier and can only be as good or bad as these initialisations are.

Deterministic Annealing (DA), proposed in [2], can be used for a variety of tasks, e.g. compression, pattern recognition, classification and can also be applied to spectral clustering of SAR images.
DA uses a data-specific distance measure. Any proper distance could be used here, e.g. those mentioned above.
Instead of hard clustering DA works with association probabilities and does not need any kind of handmade initialisation, which can be hard to obtain and lead to initialisation dependent clustering results.

2 Wishart based distance measures

Under the assumption of reciprocity of natural targets a fully-polarimetric SAR data point can be represented as a vector $\vec{\mathbf{s}}$ by

$$\vec{\mathbf{s}} = (S_{HH}, \sqrt{2}S_{HV}, S_{VV})$$

where S_{RT} is a complex component of the scattering matrix and $R \in \{H, V\}$ is the receive and $T \in \{H, V\}$ is the transmit polarisation.
In order to reduce speckle and to get better second order statistical properties, the data are often represented by the spatially averaged sample covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} \vec{\mathbf{s}}_i \vec{\mathbf{s}}_i^H$$

where H denotes the Hermitian transpose and n is the number of samples used for averaging.
If the distribution of $\vec{\mathbf{s}}$ is a multivariate complex Gaussian

with zero mean, which is a standard assumption when dealing with fully-polarimetric SAR data, the sample covariance matrix $\mathbf{C}$ of $\vec{\mathbf{s}}$ is complex Wishart distributed.

$$\vec{\mathbf{s}} \sim N(0, \mathbf{\Sigma}) \Rightarrow \mathbf{C} \sim W(n, \mathbf{\Sigma})$$

Hence, the density of $\mathbf{C}$ given the true covariance matrix $\mathbf{\Sigma}$ is defined by

$$p_n(\mathbf{C}|\mathbf{\Sigma}) = \frac{n^{nq}|\mathbf{C}|^{n-q}\exp(-n \cdot tr(\mathbf{\Sigma}^{-1}\mathbf{C}))}{|\mathbf{\Sigma}|^n \cdot \pi^{q(q-1)/2}\prod_{k=1}^{q}\Gamma(n-k+1)},$$

where $|\cdot|$ is determinant and $tr(\cdot)$ the trace of a matrix, $\Gamma(\cdot)$ the standard gamma function and q is the dimensionality of $\vec{\mathbf{s}}$.

In [1] a distance measure d_W was developed based on the Wishart distribution:

$$\begin{aligned} d_W(\mathbf{C}, \mathbf{\Sigma}) &= -\frac{1}{n}\ln p(\mathbf{C}|\mathbf{\Sigma}) - c \\ &= \ln(\det \mathbf{\Sigma}) + tr(\mathbf{\Sigma}^{-1} \cdot \mathbf{C}) \end{aligned}$$

Unfortunately this distance measure is not a metric, because it is neither homogeneous, nor symmetric and does not fullfill the triangle inequality.
Nonetheless this distance measure is often used and showed its effectiveness in practice. Because of this and its direct relation to the density function it will be also used in this work.

3 Deterministic Annealing

Deterministic Annealing (DA), proposed in [2], can be used for different tasks, e.g. compression, pattern recognition, classification and for clustering.
The data points are considered in their vector space without any spatial information. DA tries to divide data in this vector space into the most probable clusters. These clusters can be afterwards used to segment the image, by assigning each data point within the image with the label of the cluster it belongs to.
Unlike hard clustering algorithms, DA does not make hard decisions to which cluster a data point belongs. Instead there exists an association probability $p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})$ for each data point $\mathbf{C}^{(\alpha)}, \alpha = 1, ..., N$ (which is a complex covariance matrix in our case) and cluster center $\mathbf{Y}_i, i = 1, ..., c$ (which is matrix of complex elements, too).
If there is a high probability $p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})$, the distance $d(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i)$ between them should be small. This should hold for the whole set $\{\mathbf{C}^{(\alpha)}\}_{\alpha=1,...,N}$ of data points and set $\{\mathbf{Y}_i\}_{i=1,...,c}$ of cluster centers:

$$\begin{aligned} D(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\}) &= \sum_{\alpha=1}^{N}\sum_{i=1}^{c} d(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i) \cdot p(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i) \\ &= \sum_{\alpha=1}^{N}\sum_{i=1}^{c} d(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i) \cdot \\ &\quad p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})p(\mathbf{C}^{(\alpha)}) \\ &\stackrel{!}{=} min \end{aligned}$$

Obviously, there can be more than one distribution $p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})$, which minimize this weighted arithmetic mean of distances. According to Jaynes's maximum entropy principle (see [3] for more details), the distribution with the highest entropy should be choosen. Using another distribution means to assume a level of certainty, that is not justified by the given data. To choose the distribution with highest uncertainty - that with highest entropy - ensures to use only as much certainty as the data provides.
The entropy H of two sets of samples of two random variables is given by

$$\begin{aligned} H(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\}) &= -\sum_{\alpha=1}^{N}\sum_{i=1}^{c} p(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i) \cdot \\ &\quad \log(p(\mathbf{C}^{(\alpha)}, \mathbf{Y}_i)) \end{aligned}$$

but can be decomposed to

$$H(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\}) = H(\{\mathbf{C}^{(\alpha)}\}) + H(\{\mathbf{Y}_i\}|\{\mathbf{C}^{(\alpha)}\})$$

where

$$\begin{aligned} H(\{\mathbf{C}^{(\alpha)}\}) &= -\sum_{\alpha=1}^{N} \mathbf{C}^{(\alpha)} \log(p(\mathbf{C}^{(\alpha)})) \\ H(\{\mathbf{Y}_i\}|\{\mathbf{C}^{(\alpha)}\}) &= -\sum_{\alpha=1}^{N} p(\mathbf{C}^{(\alpha)}) \cdot \\ &\quad \sum_{i=1}^{c} p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)}) \log(p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})) \end{aligned}$$

The source entropy $H(\{\mathbf{C}^{(\alpha)}\})$ is independent of clustering and can therefor be dropped in further calculations.
The optimization problem can now be reformulated in terms of the distances $D(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\})$ and the conditional entropy $H(\{\mathbf{Y}_i\}|\{\mathbf{C}^{(\alpha)}\})$:

$$\begin{aligned} F(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\}) &= D(\{\mathbf{C}^{(\alpha)}\}, \{\mathbf{Y}_i\}) \\ &\quad -T \cdot H(\{\mathbf{Y}_i\}|\{\mathbf{C}^{(\alpha)}\}) \end{aligned}$$

The goal is to minimize this cost function with respect to the association probabilities $p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})$ and cluster centers $\mathbf{Y}_i$.
T is a weighting factor, which controls the tradeoff between maximizing the entropy and minimizing the distances. It can be considered as a temperature parameter

and will start at a large value, so at the beginning the entropy is maximized. When T is lowered there will be a tradeoff between maximizing H and minimizing D. For very low T hard clustering emerges.

Within the DA framework there are c cluster centers $\mathbf{Y}_i$, which partion the input space by the association probabilities $p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})$.
An essential question in any clustering approach is by how many clusters the data should be represented. There cannot be a satisfactory answer in general, because it depends on the user respectively the application how detailed one wants to represent the data. Most unsupervised clustering approaches define a certain number of clusters beforehand and initialize them either by random or by a fast and simple method. DA starts with only one cluster and creates new clusters if needed until a user-defined maximal number is reached.
A great advantage of DA is that there is no need for any kind of handmade initialisation. Consider an infinite number of cluster centers, named code vectors. Because the temperature T is very high at the beginning, the structural properties of the data plays a minor role and only the entropy is maximized. This means that all data points belong to all clusters with the same probability. That is why all code vectors lie in the center of gravity and form one big effective cluster. Be aware, that the general term "cluster" has two different meanings here: On the one hand there are the theoretical infinite number of cluster centers named code vectors, but because of the temperature parameter most of them will be at the same place and form a finite number of effective clusters. To avoid misunderstandings the infinte number of cluster centers will be called code vectors from now on and the center of the effective clusters will be termed in short as cluster centers $\mathbf{Y}_i$, where $p(\mathbf{Y}_i)$ will be the fraction of code vectors within effective cluster $\mathbf{Y}_i$.
Although there are infinte codevectors, there will only be a finite number of cluster centers in each time step, in particular only one at high temperature. Although dealing with an infinite number of code vectors only the effective clusters have to be observed.
As mentioned before the temperature will be decreased during annealing and the code vectors will be perturbed by a small value. If the temperature is still too high every code vector will drift back to the centroid of the current cluster. However, once a critical temperature is reached the code vectors will drift apart and the clusters split in distinct clusters.For a sufficient low temperature there would be as many clusters as data points. That is why the user had to set a maximum number of clusters. Creation of new clusters will be performed only if the current number of clusters is less than the maximum number. However, the annealing process will be continued until a lower bound of the temperature is reached.
The association probabilities as well as the cluster centers will be calculated in an iterative manner.

Minimizing $F(\{\mathbf{C}^{(\alpha)}\},\{\mathbf{Y}_i\})$ with respect to the association probabilities and cluster centers is straight forward. Following [2] the solution under the so called mass constraint (avoids the disadvantage to be dependent on the number of code vectors describing the effective cluster) is given by the following equations:

$$\mathbf{Y}_i = \frac{\sum_{\alpha=1}^{N} \mathbf{C}^{(\alpha)} p(\mathbf{C}^{(\alpha)}) p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)})}{p(\mathbf{Y}_i)} \quad (1)$$

$$p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)}) = \frac{p(\mathbf{Y}_i)\exp(-d(\mathbf{C}^{(\alpha)},\mathbf{Y}_i)/T)}{\sum_{j=1}^{c} p(\mathbf{Y}_j)\exp(-d(\mathbf{C}^{(\alpha)},\mathbf{Y}_j)/T)} \quad (2)$$

$$p(\mathbf{Y}_i) = \sum_{\alpha=1}^{N} p(\mathbf{C}^{(\alpha)}) p(\mathbf{Y}_i|\mathbf{C}^{(\alpha)}) \quad (3)$$

If one uses the euclidian distance in (2) the critical temperatures, where one cluster splits into more clusters, can be calculated beforehand. However, such a solution could not be obtained so easy for Wishart based distance measures. That is why each cluster center is represented by two code vectors. They are perturbed a little bit by changing the association probabilities with a sufficiently large but small enough value that stability is guaranteed. They will drift back if the temperature is too high and fall apart if critical temperature is reached. Then two new clusters will be created at the place of the two distinct code vectors.

As mentioned before one advantage of DA is that - in contrary to many other clustering algorithms - there is no need for a user-defined initialisation. Furthermore, for every temperature there exist only as many clusters as needed. DA uses benefits of annealing (e.g. avoids many poor local minima) but, because of its deterministic nature, it is much less computational expensive than simulated annealing.

4 Results

The image at the left of figure 1 is the color coded (and scaled) representation of a 347×1106 fully-polarimetric SAR image. It was speckle-filterd by a simple boxcar filter beforehand.
The clustering result, shown at the right side of figure 1, was obtained with DA under use of d_W and consists of 10 spectral clusters.

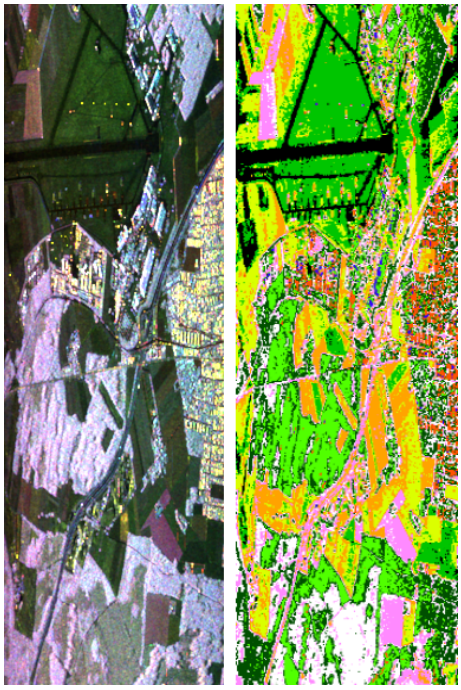

Figure 1: fully-polarimetric SAR data (E-SAR, L-Band, Oberpfaffenhofen) and DA clustering result (10 cluster)

Figure 2 shows the segmentation in more detail. The left side shows the original image data while at the right side the segmentation result is displayed. The algorithm is able to distinguish between different kinds of landuse. Although no spatial information is used, the created image segments are coincident with spatial structures in the image.

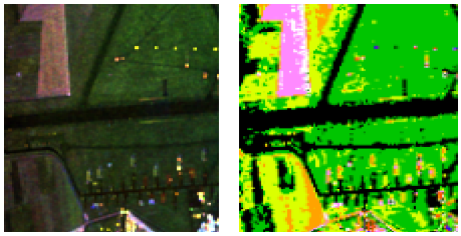

Figure 2: left: original; right: segmenatation

The clustering result of DA was compared with results of other aproaches in terms of data-log-likelihood L:

$$
\begin{aligned}
L_p(\{\mathbf{C}^{(\alpha)}, \mathbf{Y}^{(\alpha)}\}) &= \frac{1}{N} \log p(\{\mathbf{C}^{(\alpha)}\}|\{\mathbf{Y}^{(\alpha)}\}) \\
&= \frac{1}{N} \log \prod_{\alpha=1}^{N} p(\mathbf{C}^{(\alpha)}|\mathbf{Y}^{(\alpha)}) \\
&= \frac{1}{N} \sum_{\alpha=1}^{N} \log p(\mathbf{C}^{(\alpha)}|\mathbf{Y}^{(\alpha)})
\end{aligned}
$$

where $\mathbf{Y}^{(\alpha)}$ is the center of that cluster, to which the algorithm assigned the data point $\mathbf{C}^{(\alpha)}$ and p is the Wishart distribution. Therefore the same image data was clustered with the RAT [4] implementations of [5, 6]. As can be seen in figure 3 DA is superior to these standard clustering methods.

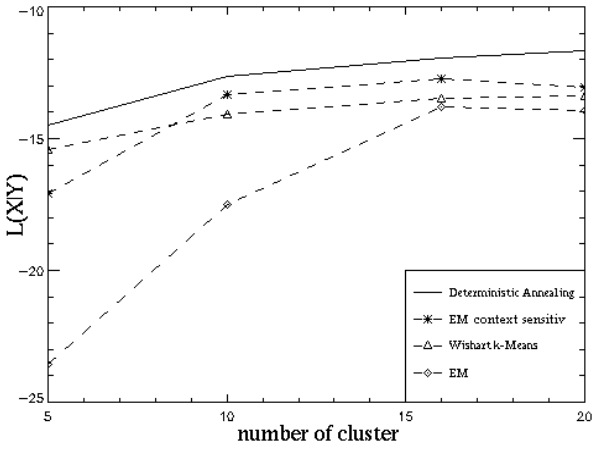

Figure 3: data log-likelihood

5 Conclusion

Deterministic annealing, which is known to have better properties than standard clustering methods, was used to segment a fully-polarimetric SAR image. The results of this method do not depend on a proper initialisation like other segmentation algorithms for SAR data. The obtained segmentation results are very promising and superior to those of previous segmentation algorithms in terms of data log likelihood for fully-polarimetric SAR data.

References

[1] J.-S. Lee, M.R. Grunes, R. Kwok: *Classification of multilook polarimetric SAR imagery based on complex Wishart distribution*, Int. J. Remote Sens., vol. 15, pp. 2299-2311, 1994.

[2] K. Rose: *Deterministic Annealing for Clustering, Compression, Classification, Regression and Related Optimization Problems*, Proc. IEEE, vol. 86, pp. 2,210-2,239, 1998.

[3] E.T. Jaynes: *Information theory and statistical mechanics*, The Physical Review. 106, Nr. 4, 15. Mai 1957, S. 620-630.

[4] http://www.cv.tu-berlin.de/rat/

[5] J.-S. Lee et all: *Unsupervised Classification Using Polarimetric Decomposition and the Complex Wishart Classifier*, EEE Trans. Geosci. Remote Sens., vol.37, no.5, pp.2249-2258, Sept. 1999.

[6] A. Reigber, M. Jäger, M. Neumann, L. Ferro-Famil: *Polarimetric fuzzy k-Means classification with consideration of spatial context*, Proceedings of POLINSAR'07, Frascati, Italy, January 2007.