Onboard Detection of Snow, Ice, Clouds and Other Geophysical Processes Using Kernel Methods

Ashok N. Srivastava ASHOK@EMAIL.ARC.NASA.GOV
Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Field, CA 94035 USA

Julienne Stroeve STROEVE@KODIAK.COLORADO.EDU
National Snow and Ice Data Center, University of Colorado at Boulder, Boulder, CO 80309 USA

To be published in the Proceedings of the ICML 2003 Workshop on Machine Learning Technologies for Autonomous Space Sciences

Abstract

The detection of clouds within a satellite image is essential for retrieving surface geophysical parameters from optical and thermal imagery. Even a small percentage of cloud cover within a radiometer pixel can adversely affect the determination of surface variables such as albedo and temperature. Thus, onboard processing of satellite data requires reliable automated cloud detection algorithms that are applicable to a wide range of surface types. Unfortunately cloud-detection, particularly over snow- and ice-covered surfaces, is a problem that plagues the field of remote sensing because of the lack of spectral contrast. This paper discusses preliminary results based on *kernel methods* for unsupervised discovery of snow, ice, clouds, and other geophysical processes based on data from the MODIS instrument and discusses implementation in computationally constrained environments such as those found on satellites.

1. Motivation for Snow, Ice, and Cloud Detection

Common approaches of detecting cloud cover are based on spectral contrast, radiance spatial contrast, radiance temporal contrast, or a combination of these methods. These techniques work well over dark targets (e.g. vegetation), since clouds appear brighter (higher albedo) in the visible range, and have lower temperatures in the infrared compared to the cloud-free background. Threshold values are then chosen to represent the cloud-free background. Problems with this method however, are that thresholds typically need to be selected from scene to scene. Another type of cloud detection that does not require absolute thresholds evaluates the spatial coherence of the observed scene. However, coherence tests suffer from the fact that false detection is likely for clear pixels directly adjacent to cloud pixels. Recent work by Stroeve (2002) shows that the current cloud masking procedure used in the Advanced Very High Resolution Radiometer (AVHRR) Polar Pathfinder product is not reliable over Greenland. Because of the limitations of multispectral feature extraction from satellite imagery to adequately discriminate clouds from snow/ice-covered surfaces, artificial intelligence (AI) techniques have seen increased use for the analysis of remotely sensed data. Other techniques, such as using a paired-histogram approach, have been attempted for estimating clouds from snow- and ice-covered surfaces. Results showed that in the polar regions, regional classifiers provide somewhat higher classification accuracy but that the algorithm still had difficulty discriminating between snow/ice and cirrus clouds. Key et al. (1999) used both a neural network approach and a traditional maximum likelihood method for cloud classification in the Arctic. They found that the neural network has greater flexibility to classify indistinct classes. Maximum likelihood results could be made to agree more closely with manual interpretation if the training areas were significantly expanded. However, at the time of the study, such a degree of training was found to be impractical for remote sensing studies because of the volume of imagery that had to be processed.

Onboard algorithms for discovery of geophysical processes such as snow, ice, and clouds have the further constraint that they need to be able to operate with high computational efficiency. Thus, in designing such discovery algorithms, one is faced with the trade-off between computational complexity (as represented by peak CPU and RAM requirements) and

accuracy. Kernel methods with appropriate modifications, may offer a computationally viable means to address this knowledge discovery problem. Preliminary results discussed here and elsewhere (Lee et al., 2003) in applying kernel methods to this problem domain indicate promising results with respect to the ability to discriminate between underlying features. Further research needs to be done to make these algorithms fully functional in an onboard setting.

Currently, most cloud detection algorithms operate in ground-based data centers. Our motivation for creating onboard discovery algorithms is to allow for rapid capture of images that include interesting phenomenon, and to reduce transmission of obscured images. Furthermore, discovery algorithms that assign pixels to equivalence classes or cluster centers can significantly increase the compressibility of the data set, thus increasing the amount of information throughput. Thieler and Gisler, (1997) showed that the number of bits per pixel saved for a *lossless* compression scheme is:

$$B = \frac{d}{2} log_2(\frac{V_0}{V}) - S \quad (1)$$

where d is the number of spectral channels, V_0 is the within cluster point scatter of the full data set (which is independent of the clustering), V is the average within cluster scatter, and S is the Shannon entropy (in bits) of the clustering: $S = -\sum_k p_k log_2 p_k$, where p_k is the fraction of the number of points that falls in cluster k. We estimate that using our kernel clustering method described below, we will save approximately 7 bits per pixel. The kernel clustering described here does not optimize the clustering for maximal compression. Theiler and Gisler, (1997) describe a method based on contiguity enhanced k-means clustering that can be adapted to optimize compression. If one allows for lossy compression, the increase in throughput is even more dramatic, since each multispectral vector would be replaced by a single number.

2. Kernel Methods

We begin by giving a brief introduction to kernel methods and our model of hyperspectral data. The kernel methods discussed here have little relation to the established notions of kernel density estimation (such as mixtures of Gaussians, Parzen windows, etc). We model the hyperspectral data as a spatiotemporal random function $Z_t(\alpha, \beta, \lambda)$, which represents a series of length T of three dimensional data cubes of size $(n \times n \times \Lambda)$, where n denotes the number of pixels in one direction (assuming square images, without loss of generality), Λ denotes the total number of measured wavelengths, and T denotes the total number of time samples.

Kernel functions can be interpreted as a similarity measure that can be general or tailored to the specific domain from which the data arises. A common measure of similarity between two spectral wavelengths λ and λ' at a given time τ_0 is captured in the linear covariance function. Linear covariance can be generalized by introducing a highly nonlinear function Φ that maps data from the Λ dimensional space to a high (possibly infinite) dimensional Hilbert space (Cristianini & Shawe-Taylor, 2000): $\Phi : \mathcal{R}^\Lambda \mapsto \mathcal{H}$. The image space of the mapping function is also known as the feature space. Thus, we can write the covariance matrix in terms of the mapped data as follows:

$$\begin{aligned} \mathbf{\Sigma}^{\Phi}_{\tau_0}(\lambda, \lambda') &= Cov(\Phi(Z_{\tau_0}(\mathbf{u}, \lambda)), \Phi(Z_{\tau_0}(\mathbf{u}, \lambda'))) \\ &= \frac{1}{N} \sum_{i=1}^{N} [\Phi(Z_{\tau_0}(\mathbf{u}_i, \lambda)) - m^{\Phi}_{\tau_0}(\lambda)] \times \\ &\quad [\Phi(Z_{\tau_0}(\mathbf{u}_i, \lambda')) - m^{\Phi}_{\tau_0}(\lambda')]^T \end{aligned}$$

$m_{\tau_0}(\lambda)$ is the mean spectral energy at wavelength λ at time τ_0 and $\mathbf{u}_i$ is the ith spatial coordinate vector, and $N = n^2$. Once a mapping Φ is prescribed, one can perform *linear* operations in the feature space and map the results back to the original Λ-dimensional space (Cristianini & Shawe-Taylor, 2000). Generic kernels include the Gaussian kernel, $K(Z_i, Z_j) = \exp - \frac{||Z_i - Z_j||^2}{\sigma^2}$ and the polynomial kernel, $K(Z_i, Z_j) =< Z_i, Z_j >^p$. Note that we recover the linear covariance measure if we take $p = 1$ in the polynomial kernel.

The MODIS data that we currently have is not labelled, so rather than casting this as a classification problem, we formulate it as a kernel clustering problem as shown in (Girolami, 2001). This raises the possibility that other geophysical processes in the data could be discovered. An advantage of the kernel approach is that once a suitable kernel is obtained, it can be applied to both clustering and classification problems. A kernel function can be uploaded to a satellite for onboard science in a clustering or classification setting, depending on the nature of the mission. In a deep space probe, for instance, it is likely that it would be used for both clustering and classification: clustering to generate features and then classification to predict those features in new images.

3. Kernel Clustering in Feature Space

Girolami, (2001) has given an algorithm to perform clustering in the feature space using an approach similar to k-means clustering. A brief review of the method follows. The cost equation for k-means clustering in

the feature space at a given instant in time is:

$$G^{\Phi} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ki}[\Phi(Z_i) - m_k^{\Phi}]^T[\Phi(Z_i) - m_k^{\Phi}] \quad (2)$$

where q_{ki} is the cluster membership indicator function ($q_{ki} = 1$ if vector Z_i is a member of cluster k, and zero otherwise), and m_k^{Φ} is the cluster center in the feature space. Thus, if we expand the right-hand side of the above equation, and take $m_k^{\Phi} = \frac{1}{N}\sum_{i=1}^{N} q_{ki}\Phi(Z_i)$, which represents the centroid of the cluster in feature space, we obtain an equation in which only inner products appear. The nonlinear mapping Φ does not need to be determined explicitly because the kernel function is taken as the inner product in the feature space: $K_{ij} = \Phi^T(Z_i)\Phi(Z_j)$. The objective of kernel clustering is to find a membership function q and cluster centers m^{Φ} that minimize the cost G^{Φ}. Various methods can be used to minimize G^{Φ}, including annealing methods (as described in Girolami, 2001) or direct search.

4. Designing Probabilistic Kernels

Although there are many generic positive definite kernel functions such as the Gaussian kernel, the polynomial kernel, and the Fourier kernel, our goal is to create kernels that have superior performance with respect to the discovery process compared to a generic kernel and are robust to errors made in the underlying assumptions about the distribution of the data. In order to accomplish this, we build a variant of Probabilistic Kernels (P-Kernels) based on the data, which is an idea that is discussed in (Haussler, 1999). The underpinnings of the P-Kernel is a probabilistic model of the data. In our case, we build a Gaussian mixture model with spherical covariance structure: $P(\mathbf{Z}|\theta) = \sum_{k=1}^{K}\beta_k G(\mathbf{Z}|\theta_k)$, where $\theta_k = (\mu_k, \mathbf{\Sigma}_k)$, and G is the multivariate normal distribution.

Haussler's P-Kernel takes the kernel function to be $K(Z_i, Z_j) = P(Z_i|\mathcal{M})P(Z_j|\mathcal{M})$, where $\mathcal{M}$ is a probabilistic model. This kernel assumes that the observations are independent and is very sensitive to small changes in the uncertainty of the probabilities. Our goal is to build a kernel that is insensitive to variations in the estimated probability distribution. We thus design the following new Probabilistic Kernel:

$$K^P(Z_i, Z_j) = \left[\frac{\sum_{k=1}^{K} P(k|Z_i)P(k|Z_j)}{||P_i||||P_j||}\right]^2 \quad (3)$$

where K^P denotes the Probabilistic Kernel, K is the number of modes in the mixture model, $P(k|Z_i)$ is the posterior probability of spectrum Z_i being allocated to class k, and $||P_i||$ is the norm of the vector of class distributions for spectrum i. This kernel treats uncertainty in a very different way than the standard P-kernel. If there is no uncertainty in the class distribution, the distribution will have a delta function at a single value of k, thus the kernel function will either be zero if Z_i and Z_j belong to different classes, or one if they belong to the same class. In this model, if there is uncertainty in the class distribution, the entropy of the class distribution will be closer to a maximum (i.e., the distribution will be less peaked), thus affecting the inner product in equation 3 and giving intermediate values between zero and one for the kernel function. This feature becomes apparent when looking at experimental results of multispectral data taken over Greenland. In our experiments, we used this new Probabilistic Kernel in the kernel clustering algorithm.

5. Data and Experimental Results

We obtained MODIS level 1B data for the Greenland ice sheet from the NASA Langley DAAC and mapped the data to a 1.25 km equal-area scalable Earth-grid (EASE-grid) using software developed by NSIDC to process MODIS level 1B data and convert the visible channel data to top-of-the-atmosphere (TOA) reflectances. We chose to work with Greenland images because the second author has performed extensive fieldwork in the area. Next the TOA reflectances were normalized by the cosine of the solar zenith angle. Only the first 7 MODIS channels were used for this study [1]. The image shown in Figure 1 was taken on June 1, 2000 at 1445 GMT. As expected, the spectral signals for the 7 different MODIS channels are highly correlated, with linear correlation coefficients over 98%.

Figure 1a shows the TOA reflectance corresponding to MODIS channel 5. The darker areas correspond to regions of lower reflectance than the brighter areas, indicating regions of larger snow grain size. Analysis of MODIS Channel 1 data from June 1, 2000 indicate that most of the ice sheet is snow covered. Exceptions occur along the margins of the ice sheet in the south where there is either bare ice or bare ground exposed. These areas are evident by the even darker (lower reflectance) regions shown in Figure 1a. Some of the dark areas in the south west correspond to larger snow

[1] The bandwidths of the first seven MODIS channels are as follows (in nm): Channel 1: 620-670, Channel 2: 841-876, Channel 3: 459-479, Channel 4: 545-565, Channel 5: 1230-1250, Channel 6: 1628-1652, Channel 7: 2105-2155. Resolutions for Channels 1-2: 250 m, 3-7: 500 m

grain sizes associated with melt processes. Analysis of passive microwave melt data show that some melt is occurring in this region on June 1, 2000.

To establish a baseline of performance, we performed clustering on the data using the k-means clustering algorithm with 10 centers specified. The algorithm was seeded with random initial cluster centers. In order to encode some notion of spatial coherence, we vectorized 5x5 blocks of the data and performed k-means on the resulting $5 \times 5 \times 7 = 175$ dimensional vectors. There are other methods to encode spatial coherence based on variants of the EM algorithm (Masson & Pieczynski, 1993) that bias the model towards a smoother spatial representation. Results for a typical run are shown in Figure 1b, where each distinct level of gray corresponds to a different cluster center. The algorithm captures some of the clouds over water regions and classifies Greenland as a single, monolithic entity.

We ran the kernel clustering algorithm using the Gaussian kernel function with 10 centers using the vectorized data in the full 175 dimensional space. The Gaussian kernel has an isotropic scale parameter which requires tuning. We chose to set the scale parameter as the average minimum Euclidean distance between the centroids discovered in the k-means algorithm (Cristianini & Shawe-Taylor, 2000). These results are shown in Figure 2a. The results have approximately a 50% overlap with the k-means results. There is a marked difference in the algorithm's performance over the Greenland ice sheet. Rather than attributing a single cluster to the entire region, the kernel method breaks the region into two areas. Upon further investigation, we determined that the algorithm revealed regions of different snow types (e.g. different grain sizes). This is of physical relevance since grain size directly affects the calculation of albedo.

To determine whether the results of the k-means clustering were statistically different than the kernel method, we ran ten models of kernel clustering and k-means clustering with random initial conditions and recorded the number of cluster centers that were used to describe the Greenland ice sheet using each algorithm. A one-way analysis of variance indicated that we must reject the null hypothesis that the means of the number of cluster centers used to describe the ice sheet are equal at the 0.1% significance level with $F = 15.78$ and $p = 0.0009$. The average number of cluster centers used to describe the ice sheet is 1.5 for k-means clustering and 2.6 for kernel clustering. Figure 2b shows the results of the application of the Gaussian Mixture Model (GMM) to the same data set. Unlike the k-means segmentation, the GMM discovers

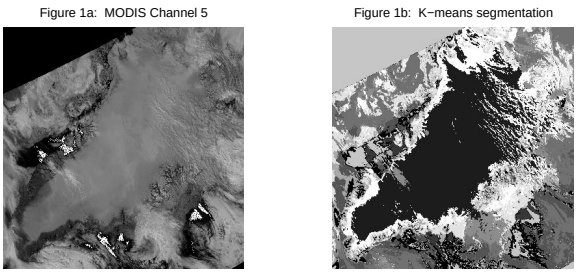

Figure 1. **(a)** This figure shows the TOA reflectance of MODIS channel 5 (1230-1250 nm). The darker areas over Greenland are lower reflectance than lighter regions, and correspond to larger snow grain sizes, which according to passive microwave melt estimates on that day could indicate some melting in the south western margin of the ice sheet. **(b)** Shows the k-means clustering results. Notice that the entire ice sheet is characterized as a single entity.

different snow types (due to the difference in grain size) on the ice sheet and also discovers water regions (lighter color in lower left of Figure 2b). As in the case of the Gaussian Kernel and the k-means algorithm, the GMM isolates clouds over the ice sheet (NE and SW regions). Notice that the gradient in grain size, which indicates a region in flux, is not detected.

Figure 2c shows the output of the Probabilistic Kernel model described above. As expected, the output is similar to the GMM, with the exception of the identification of the gradient in grain size around the ice sheet, particularly in the SE region. Figure 2d shows the difference between the Probabilistic Kernel and the GMM. Another area that the two models differ is in the identification of water regions. The GMM isolates water regions whereas the Probabilistic Kernel model is combining those regions with other regions of low reflectance. Overall the models differ in about 4.4% of the predictions. The Probabilistic Kernel method isolates regions of uncertainty and attributes them to a given class, sometimes distinct from the underlying estimated probability distribution, in an attempt to enhance the signal to noise ratio. We are actively pursuing enhancing the kernel model with domain knowledge to further increase its value.

6. Challenges of Applying Kernel Methods in Onboard Processing

While kernel methods offer the promise to improve clustering results, they suffer from some drawbacks that make their use in onboard applications challenging. Kernel methods can be computationally and memory intensive because they require computing and storing an $(N \times N)$ kernel matrix, where N is the number of independent data points. An important area of

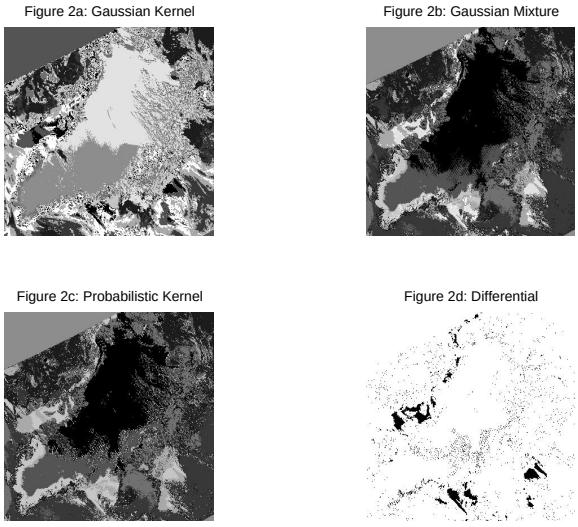

Figure 2. **(a)** The results of applying kernel clustering to MODIS data from June 2000 using the Gaussian kernel. The kernel method discovers the clouds over snow (NE region), plain snow (S part of ice sheet), clouds over water (SE region), and also divides Greenland into two areas which have different grain sizes. Since ground-truth data is not available, we provide an image from Channel 5 in Figure 1a which is used to detect grain size to corroborate our results. This result motivates our attempt to design kernels to discover geophysical processes. **(b)** The results of application of a Gaussian Mixture Model. This isolates the two regions with different grain size (with a different and more accurate boundary) compared with the Gaussian kernel. **(c)** Application of the probabilistic kernel reveals the same area with different grain size but also indicates the gradient in grain size. **(d)** The difference between the GMM and the Probabilistic Kernel clustering output.

research is in regards to reducing the memory requirements of kernel methods.

Other issues that arise in designed kernels is that they may not appropriately express the inherent similarity in the data. For example, it may be that a designed similarity metric is not as robust as a Euclidean metric, even though it encodes other domain knowledge. The Euclidean metric for hyperspectral data has well understood behavior if one controls for outliers. Other metrics, such as linear correlation and Mahalanobis distance have similar properties.

The preliminary results presented here lend credibility to the idea that a kernel-based method could reveal geophysical processes and discriminate between such processes. The Probabilistic Kernel may be indicative of a new method to generate kernels directly from data which appropriately deal with uncertainty in the underlying probability distribution. Onboard algorithms based on such a kernel could dramatically increase the transmission of useful data while reducing the transmission of obscured images.

Acknowledgements

The authors would like to thank the reviewers, Bill Macready, Nikunj Oza, and Brett Zane-Ulman for valuable comments and suggestions. This work was supported by the NASA Computing, Information, and Communication Technology (CICT) program and the Intelligent Data Understanding project.

References

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press.

Girolami, M. (2001). Mercer kernel based clustering in feature space. *IEEE Transactions on Neural Networks*, *13*, 780–784.

Haussler, D. (1999). *Convolution kernels on discrete structures* (Technical Report). University of California Santa Cruz.

Key, J. (1999). The cloud and surface parameter retrieval (caspr) system for polar avhrr. *Coop. Inst. for Meteorol. Satellite Stud.*

Lee, Y., Wahba, G., & Ackerman, S. A. (2003). *Cloud classification of satellite radiance data by multicategory support vector machines* (Technical Report). University of Wisconsin.

Masson, P., & Pieczynski, W. (1993). SEM algorithm and unsupervised statistical segmentation of satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, *31*, 618–633.

Stroeve, J. (2002). Assessment of greenland albedo variability from the avhrr polar pathfinder data set. *Journal of Geophysical Research*, *33*, 989–1034.

Theiler, J., & Gisler, G. (1997). A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation. *Proceedings of SPIE (International Society for Optical Engineering)* (pp. 108–118).