ON THE EVALUATION OF SYNTHETIC HYPERSPECTRAL IMAGERY

Michael J. Mendenhall

Dept of Electrical and Computer Engineering
Air Force Inst of Technology, WPAFB, OH

Erzsébet Merényi

Dept of Electrical and Computer Engineering
Rice University, Houston, TX

ABSTRACT

In developing algorithms that exploit model-generated data, it is important to understand the realism of the data generated by that model. One way to address this issue is to exercise a well understood, yet diverse process, that will help draw out the strengths and weaknesses of the data generation system. We accomplish this by using a typical chain of processing steps on a synthetic hyperspectral image created by the Digital Imaging Remote Sensing Image Generation (DIRSIG) tool [1]. The clustering, classification, and feature selection, which are part of this processing, are used to assess the realism of the data based on the performance compared to previous similar analysis on real hyperspectral data.

Index Terms— Self-organizing map, relevance learning, learning vector quantization, synthetic hyperspectral imagery

1. INTRODUCTION

Hyperspectral imagery has found widespread use in commercial and defense applications. The complexity of the problem domains and datasets creates a need for sophisticated algorithms. These algorithms require validation, which is costly as scene truthing involves non-trivial logistics, and access to samples and the accuracy of the sampling is often limited.

One potential alternative to more traditional scenario development, data acquisition, and ground truthing is to use synthetic imagery from a system such as the Digital Imaging Remote Sensing Image Generation (DIRSIG) tool [1]. Synthetic data allow for flexible scenario development, reduced acquisition cost, and the availability of absolute ground truth. The importance of absolute knowledge of ground truth should not be undersold; it enables *accurate* algorithm assessment provided the image synthesis process is realistic. By realistic we mean that the data generation system should not make the synthetic image so pristine that it does not reflect noise sources such as those present in the sensor or environment. The synthesis should also avoid imposing some specific "well-behaved" statistical distribution on the data in the case of hyperspectral imagery since it is a widely observed fact that statistical models are hard to find for such data. Analysis on data that are too pristine can lead to a bias in algorithm results that tend toward better performance when in fact true performance is more limited.

This article aims to analyze synthetic hyperspectral imagery from the point of view of the realism of the data exploitation results. To this end, we assess the potential of a DIRSIG synthetic hyperpectral image by emulating a typical processing chain of clustering, interpretation and labeling of clusters, followed by supervised classification, and assessment of the accuracy of the obtained thematic map.

2. EVALUATION METHODOLOGY

We use a Self-Organizing Map (SOM) [2] for clustering, and learning vector quantization (LVQ) [2] for subsequent supervised classification. Both SOM clustering and LVQ classification (which is the supervised equivalent of the SOM) have been used on real NASA/JPL AVIRIS [3] hyperspectral imagery of the Lunar Crater Volcanic Field (LCVF) image [4] [5]. SOM clustering was also used previously on an AVIRIS image of Ocean City [6]. In both cases, performance is well understood and is used as a comparison for the work here. In all steps of this processing chain, we compare the outcome to the quality — level of detail, accuracy, and similarity/differences — of the results we obtained in previous studies on real data.

Comparison of the results of clustering may be semi-qualitative, whereas evaluation of classification results is purely quantitative. However, in both cases, the results of the process are described in a manner that is consistent with the interpretation of previously obtained results.

We use a SOM for clustering because the SOM is capable of discovering many clusters with a wide variety of shapes, sizes, densities, and proximities, in contrast to many other techniques which are sensitive to the data distribution. For example, K-means favors hyperspherical clusters and may not capture other shapes accurately. In hyperspectral dimensions, the structure of the data space can be quite complicated in terms of the variability of the statistical properties of the similarity groups, which is exacerbated by the potentially large number of the clusters. In [6] each of 38 SOM

MM: The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

EM thanks NASA AISR grant support NNG05GA94G

clusters could be associated with a unique physical entity in the scene. In contrast, ISODATA (K-means) produced very confused results from the same data.

For background on SOMs, see [2]. Briefly, it is an unsupervised neural learning algorithm that maps an n-dimensional data manifold $M \subset \mathbb{R}^n$ to prototype (weight) vectors attached to neural units, and organizes the prototypes in a lower (1- or 2-) dimensional rigid lattice A of N neural units. The weight vector w_i of each neural unit i is adapted iteratively by repeated application of the following steps: Find the best matching unit i for data vector $v \in M$, such that

$$\|v - w_i\| \leq \|v - w_j\| \quad \forall j \in A \tag{1}$$

and update the weight w_i and its neighbors w_j according to

$$w_j(t+1) = w_j(t) + \alpha(t) h_{i,j}(t)(v - w_j(t)) \tag{2}$$

where t is time, $\alpha(t)$ is the learning rate, and $h_{i,j}(t)$ is a neighborhood function (often a Gaussian kernel) around the best matching unit w_i. Through the above process, the weight vectors become vector quantization prototypes of the input space M, distributed so as to best represent the data density. In the learned SOM, groups of similar prototypes collectively represent groups of similar data. The data points mapped to a group of similar prototypes comprise a data cluster. Our implementation is the "conscience" SOM, which produces more faithful representation of the data density than the original Kohonen SOM (see [4] and references therein).

For supervised classification we use Relevance Learning, a form of learning vector quantization [2] (LVQ), which is the supervised version of the SOM. The form of LVQ used here is based on iterative improvements to LVQ2.1 [2] and is the maximal margin classifier Generalized Relevance LVQ Improved [5] (GRLVQI). The GRLVQI is an embedded-type feature selection and classification method that optimizes feature selection for the classification problem. It has demonstrated excellent performance on real hyperspectral imagery in [5] for 7-, 23-, and 35-class problems.

In GRLVQI, the indexes of the winning prototypes are selected as:

$$c = \arg\min_q \left(\sum_{k=1}^{n} \lambda(k) \left(x_m(k) - w_q(k) \right)^2 \right), \tag{3}$$

where the Euclidean distance between input sample x_m and prototype w_q is weighted by the relevances λ. Two winners, the best matching in-class, and the best matching out-of-class prototypes, w_J and w_K, are adapted iteratively such that the winning in-class prototype is moved toward, and the out-of-class prototype is moved away, from the sample. Prototype updates are accomplished using gradient descent and are a function of the relevances in order to incorporate relevant data dimensions, for classification, in the learning process. The vector of relevance factors λ is also updated at each step by gradient descent, according to an equation coupled with the weight update which contains classification error in the cost function. Thus, directions (spectral features) that are more important for the classification will be weighted with larger relevance values. Relevances are non-negative and are scaled by their ℓ_1-norm to avoid numerical instabilities. This scaling also gives relevances a nice interpretation as probabilities.

3. SYNTHETIC HYPERSPECTRAL DATA

We use a 400 $\times$ 400 pixel image covering 800$\times$800 m^2 with 2m$\times$2m resolution, generated with the DIRSIG tool [1]. A natural color composite is shown in Fig. 1. The image was spectrally resampled to resemble a NASA/JPL AVIRIS [3] image, then atmospherically corrected using empirical line correction. Noisy bands due to water vapor absorption and other atmospheric affects were deleted resulting in 184 spectral bands. More details on this image, including description of material types in the scene, are given in [7].

4. PROCESSING CHAIN

4.1. Cluster Analysis: Determining Classes of Interest

In order to determine relevant groupings in the data, we capture groups of similar prototype vectors in the SOM by examination of the distances of the weights in data space (not in the SOM lattice). Visualization of such distances over the SOM lattice delineates boundaries of prototype (weight) clusters in the SOM. Data points mapped to a cluster of prototypes comprise a data cluster. For the prototype clusters identified in the SOM, we refer the reader (due to space limitations) to [7]. The clusters in the image are shown in Fig. 2. Many more than the indicated 38 clusters were detected by the SOM. We resorted to showing this number of clusters because it would be too confusing to use more colors on the thematic map. It is obvious, for example, that while we are color coding 15 roof types, a number of other roof types were left uncolored (small black rectangles). This image is very rich in the variety of man-made materials such as roofing materials. The 15 color coded roof types are highlighted and enlarged in a version of the cluster map that has the large background clusters (paving, light green cluster V; grass, dark green and flesh colors, K, T; and some trees) removed, in [7].

While the statistical variations of clusters are similar, the number of clusters found in the DIRSIG image using the SOM is considerably higher than that in the urban image in [6]. This may in part be a result of generating the synthetic image with a large number of roofing materials. Furthermore, the DIRSIG image has twice the spatial resolution of the urban image in [6] and little mixing within pixels.

Samples from the SOM clusters in Fig. 2 are selected, analyzed, and used in subsequent classification. Mean spectra of the samples for 34 material types, used in this study, are

Fig. 1. RGB color composite of the DIRSIG synthetic image.

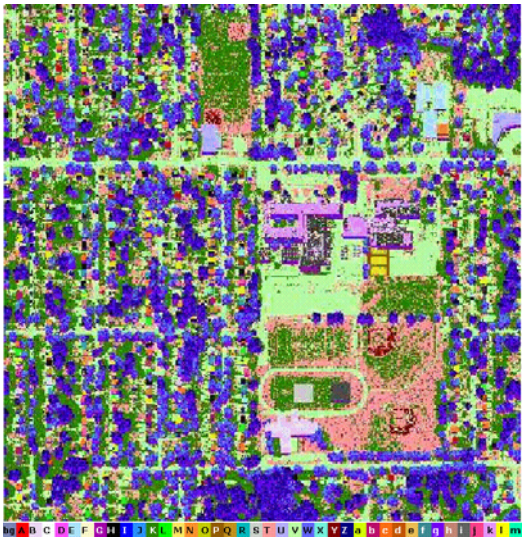

Fig. 2. Clusters found in the DIRSIG hyperspectral image.

presented in Fig. 3. Interpretation of the clusters was done by visual inspection, based on user recognition of the object functionalities (i.e., playing field, road, parking lot, tennis court, houses, trees, etc.), combined with the spectral variations among objects of similar functional categories. For example, for the different clusters delineating roof tops (e.g., A, B, F, L, N. Q, R. X, a, b, h, j, l), we verified that the spectral signatures were different, and therefore these different clusters were labeled as different classes.

Selection of training samples followed a typical situation, where a user knows parts of the scene and takes samples from those areas, for various cover types. We did this for each of the 34 selected clusters. The number of training samples taken for each class is also intended to emulate a typical remote sensing situation, where the user does not have equal access to all cover types, and therefore the obtained training samples are unevenly distributed. This is important for creating a challenging — as well as realistic — classification situation. Altogether 1811 samples were collected.

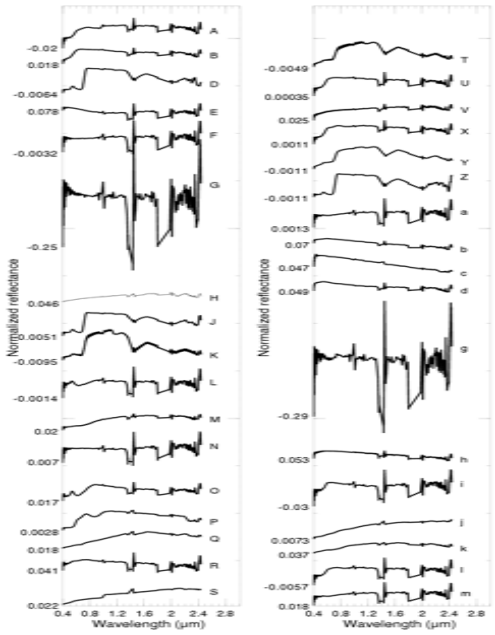

Fig. 3. Mean spectra for 34 material classes. The labels are keyed in Fig. 2.

4.2. Classification

Classification is accomplished using the GRLVQI classifier described in Section 2 with 5 prototypes per class and a similar learn schedule as in [5]. The 1811 samples across 34 material classes are used in training and testing. Classification performance is determined using 5-fold cross validation in order to balance the bias and variance of the classification and relevance factor results. Each fold has 362 (1449) testing (training) samples. Testing accuracy for each of the 5 folds is 100%. Although the accuracy is higher then anticipated, it is worth noting that the piece-wise linear decision boundary produced by GRLVQI with 5 prototypes per class on the LCVF data set is 97.2% [5], for 35-class case. The results presented show an accuracy rate of 2.8% over that of the comparison image, which is not an unreasonable outcome. Results from the LCVF data were generated using three-fold cross validation over 1464 samples, and so fewer representative samples were used in the training, which may account for some of the difference in performance. The much larger, and thus more mixed, pixels in the AVIRIS image can also work against perfect classification results. Yet another source of this difference is the physical nature of surface that is imaged. Some of the geologic materials at the LCVF site have less defined spectral boundaries than man-made materials.

Selected mean class spectra and relevances for two of the five cross-validation runs, and the average of all five runs, are presented in Fig. 4(a-c), respectively. The results here show

that a majority of the significant relevance values occur for wavelengths < 1400nm, which is consistent with the 35-class LCVF data results in [5]. There are no significant relevances for $1400 <$wavelength< 1800nm, which is different than that of the 35-class LCVF data in [5]. Finally, there are a few significant relevances for wavelength> 2000nm, which is consistent with the 35-class LCVF data in [5]. There are 40 significant relevance values greater than 0.001 for the average of the relevances over the 5 cross validation runs. (The value of 0.001 was chosen to match that used in [5].) This is nearly 63% fewer features than that reported for the 35-class LCVF problem in [5]. An interesting thing to note is that different feature combinations result in the same classification performance. This result indicates that a unique feature combination does not necessarily exist that provides best performance for a given problem. However, we note that it is possible that as the problem becomes significantly more difficult, the number of solutions that generate a given classification performance would likely diminish.

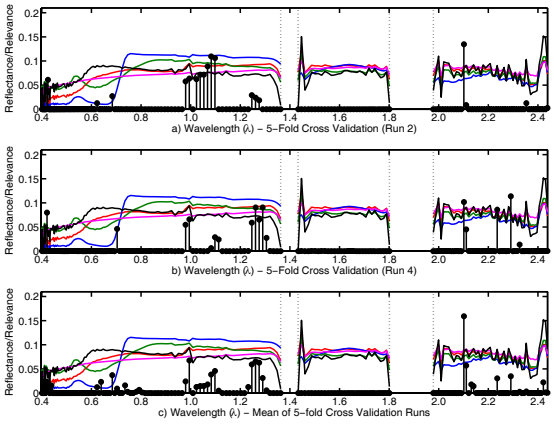

Fig. 4. Selected representative spectra (**A**-red, **J**-blue, **O**-green, **V**-magenta, **i**-black) with relevances overlaid. Areas between vertical dashed lines indicate deleted image bands due to atmospheric water absorption.

The difference in classification between the 35-class LCVF data and the 34-class DIRSIG data does not appear to be unreasonable with a performance increase of 2.8%. However, a classification accuracy of 100% for a problem of this magnitude, across all cross validation runs, does seam unlikely. The reduced set of features may be an indication that feature selection results are optimistic. The number of significant features (those with relevances greater than 0.001) is more in line with the 7-class problem reported in [5]. The difference in feature selection results may be attributed to two primary causes. First, pixels in the LCVF data set are approximately 17m$\times$17m while 2m$\times$2m in the DIRSIG image. This means that the spectra in the LCVF data set likely have more complex spectral mixtures (i.e., more material types in one pixel) then those of the DIRSIG image. Secondly, the DIRSIG image is over an urban site where man-made materials dominate and which may be more sharply distinguishable from one another than geological materials at the LCVF site (e.g., volcanic cinders and weathered cinders).

5. CONCLUSIONS

Our preliminary results suggest that synthetic hyperspectral imagery generated by DIRSIG may be viable for the development of complex exploitation algorithms. One indication of this is that the number of clusters/classes extracted from the image, and the level of detail which distinguishes the derived thematic units are very similar to those obtained in two previous analyses with the same tools, on real hyperspectral (AVIRIS) images. The seemingly unrealistic 100% classification accuracy — although only 2.8% higher than for the real LCVF data in earlier work — and the significantly lower number of spectral bands deemed relevant for the classification, are likely due to a combination of the very high spatial resolution of the DIRSIG image and crisper spectral distinctions among man-made materials than among geologic cover types. Future work will consider more complex imagery generated by DIRSIG to further examine these issues.

6. REFERENCES

[1] "The digital imaging and remote sensing image generation model," http://dirsig.cis.rit.edu/.

[2] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag Berlin Heidelberg, second edition, 2001.

[3] R. O. Green, "Summaries of the 6th Annual JPL Airborne Geoscience Workshop, 1. AVIRIS Workshop," Pasadena, CA, Mar 4–6 1996.

[4] T. Villmann, E. Merényi, and B. Hammer, "Neural maps in remote sensing image analysis," *Neural Networks*, vol. 16, pp. 389–403, 2003.

[5] Mendenhall and E. Merényi, "Relevance-based feature extraction for hyperspectral images," *IEEE Trans. on Neural Networks*, vol. 19, no. 4, pp. 658–672, April 2008.

[6] E. Merényi, B. Csató, and K. Taşdemir, "Knowledge discovery in urban environments from fused multi-dimensional imagery," in *Proc. IEEE GRSS/ISPRS Joint Workshop on Rem. Sens. and Data Fusion over Urban Areas (URBAN 2007).*, Paris, France, 11–13 April 2007, pp. 1–13.

[7] E. Merényi, K. Tasdemir, and W. Farrand, "Intelligent information extraction to aid science decision making in autonomous space exploration," in *Proceedings of DSS08 SPIE Defense and Security Symposium, Space Exploration Technologies*, W. Fink, Ed., Orlando, FL, March 17–18 2008, vol. 6960, p. 69600M, SPIE, Invited.