A New Method for Supervised Optimum Multispectral and Hyperspectral Data Subset Selection

R. K. McConnell
WAY-2C, 14 Stevens Ter., Arlington, MA, 02476, USA. (email: rkm@way2c.com)

1. Introduction

Almost as long as hyperspectral and multispectral images have been available, researchers have sought efficient methods to identify the subset of the image data most relevant for differentiation and identification of the specific classes of interest. This becomes particularly important as automated imaging systems may be called upon to control such things as robotic harvesting equipment in uncontrolled lighting conditions. Here spectra of classes of interest may change with time, environmental, and/or crop conditions.

In recent years much research has focused on information theoretic approaches involving mutual information between class and data to find the most relevant data subset. The problem has been to maximize the mutual information between class and data while minimizing redundant mutual information among the data bands selected [1-3].

We describe and illustrate a new method to quickly and efficiently accomplish this task. While often desirable, and frequently obtainable, the objective of the method is not necessarily to determine the class of each individual pixel independent of its surroundings. Sometimes it will be enough to take advantage of the exponential increase in the probability of a correct maximum likelihood classification as the number of independent pixels included in a test region increases. For many applications, particularly those involving classes with complex data distributions, the advantages of this increase can far outweigh any disadvantages of the corresponding decrease in spatial class resolution.

In addition to determining an optimum data subset, the method provides a measure of the reliability to be expected from classifications based on that subset.

2. The Data Selection Method

The method, described in more detail in [4], starts with a training cube containing labeled regions representing each of the classes of interest. The cube data may include "bands" from almost any combination of co-registered data, including such things as derived indices and possibly information from other sources. No assumptions are made about the nature of or probability distributions of the data. The main objective is to find a small subset of the data that correlates well enough with the classes of interest to provide a reliable maximum likelihood classification for user-selected regions. Once the subset is known it can be used to limit the data to be acquired, transmitted, stored and/or processed for other similar scenes.

Like several recent methods, the new method makes use of the mutual information between class and data, calculated from the training set, to determine the most relevant data. This mutual information, often referred to as relevance is given by

$$relevance = \mathrm{H}(C) + \mathrm{H}(D) - \mathrm{H}(C\&D) \qquad (1)$$

where:

H(C) is the entropy of the class probability distribution,
H(D) is the entropy of the data distribution and
H($C\&D$) is the entropy of the joint distribution of $C\&D$.

As illustrated in Fig. 1, the new method does not recognize the individual bands as the smallest units of interest. Instead it focuses on small data windows, typically containing only three adjacent bits, as the relevant data units. Thus each individual pixel in a band usually contains several different data windows. One or more of these may correlate to some significant degree with the classes of interest. Mutual information between class and data is calculated for each of the data window positions for each of the data "bands"

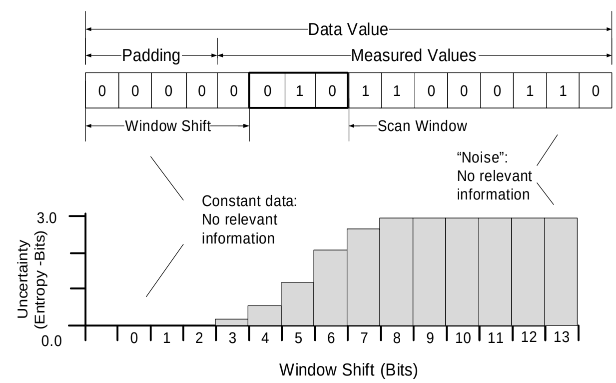

Fig. 1. Typical cube data word, scan window, and positional scan data entropy.

under consideration. The band and window within that band exhibiting the most mutual information with the class distribution is chosen as the site of the first component of most relevant data.

Once the first component of the most relevant data has been identified, the scanning process over all data windows in all data bands is repeated. However this time the data to be included in the relevance calculation is a concatenation of data from the first most relevant site with that from the scanned window. The band and window within that band contributing the most to the new mutual information value is chosen as the site of the second component of most relevant data. Using the concatenation of the data from both data sites in the relevance calculation insures that highly correlated data site pairs will not appear in the final optimum data set.

With the second component of the most relevant data identified, the scanning process over all data windows in all data bands can again be repeated. This time the data to be included in the mutual information calculation is a concatenation of data from the first most relevant site, the second most relevant site, and the scanned window. The band, and window within that band, contributing the most to the new mutual information value is chosen as the site of the third component of most relevant data. Again inclusion of data from the previous optimum units prevents highly correlated data site pairs from appearing in the final optimum data set.

In principle, the above process may be repeated indefinitely. However, practical advantages tend to keep the selected component limit to three or fewer.

Computing time is essentially linear in the volume of the cube included in the training set and the number of bands to be incorporated in the selection. The time to locate an optimum three data band combination from a 10^8 byte data cube is about a minute on a typical laptop.

A byproduct of the relevance calculation is an estimate of the probability that a single randomly chosen data value will be assigned to the correct class. We refer this as the predicted single pixel agreement or $PSPA.$ It is defined by:

$$\log(PSPA) = \mathrm{H}\left(D\right) - \mathrm{H}\left(C\,\&\,D\right) \qquad (2)$$

The value of Equ.(2) is a measure of the difference between the information necessary to specify the class and the available relevant information. As the number of selected data units (bands) increases, the number of data values required for adequate sampling does also. So there are obvious practical advantages both to keeping the number of classes of interest low, and to having data units with adequate contrast for class differentiation available. Fortunately, when modest numbers of classes must be differentiated, a combination of three or fewer spectral bands can often be found that will provide enough contrast for robust non-parametric maximum likelihood classification.

If three bands are enough to differentiate the classes then there is an additional advantage: the relevant data can be displayed in the form of a false color image compatible with viewing by humans with normal color vision. If three-bit wide data windows have been used then each data value to be included in any automated classification phase consists of only nine bits.

3. Classification Method

Since the method uses only the data bits determined to be most relevant, high order bits from the bands retained are often discarded. One result of this is that, even if the original data is characterized by a simple unimodal distribution for each class, much of the data to be analyzed is likely to have complex multimodal distributions. For such distributions non-parametric maximum likelihood classification based on comparison of observed data distributions with reference distributions for each class of interest is appropriate [5]. The training regions originally used for optimum data set selection can also be used for estimating these distributions.

Equ.(2) provides an estimate for the probability that a single pixel will be correctly classified. As mentioned above, the probability of correct maximum likelihood classification based on comparison of observed data distributions can be shown to increase exponentially with the number of independent samples. Therefore, at the expense of some spatial resolution, one can often dramatically increase the classification robustness by increasing the number of samples in the region to be classified.

4. Applications

Some suppliers of hyperspectral instrumentation may be concerned about a method that purports to be able to differentiate classes of interest based on a small number of bands. After all, one of the selling points of such instrumentation is the large number of simultaneous bands offered. We argue that the suppliers should not be worried, particularly for agricultural and similar operations where the spectra of the items of interest or the environment are variable. For such applications the advantages of multispectral and hyperspectral imaging systems may lie even more in the ability to adapt quickly to changing requirements necessary to achieve optimum contrast among classes, than in the ability to provide data for detailed broad band spectral fitting.

Even when the methods described here are employed, an imager capable of full cube capture is normally required for optimum subset determination. Such determinations can typically be made within a minute or so of complete cube capture and training region specification. Therefore, adapting data capture to obtain optimum contrast between classes of interest in near real-time becomes a possibility. If instruments allow easy selection of the data subset combination having optimum contrast, then there is the potential for robust, real-time image interpretation at tens of megapixels per second without heroic efforts.

The method has been demonstrated to yield good interpretation results for multispectral and hyperspectral images as well as a variety of non-image time-series data.

References

[1] Ren, J., Kelman, T. and Marshall, S., “Adaptive clustering of spectral components for band selection in hyperspectral imagery”, Proceedings of University of Strathclyde’s Second Annual Academic Hyperspectral Imaging Conference, 17-18 May 2011.

[2] Guo, B., Gunn, S. R., Damper, R. I. and Nelson, J. D. B., “Band selection for hyperspectral image classification using mutual information,” IEEE Geoscience and Remote Sensing Letters, 3(4): 522-526 (2006).

[3] Sotoca, J. M., Pla, F. and Sanchez, J. S., “Band selection in multispectral images by minimization of dependent information”, IEEE Trans. Systems, Man, and Cybernetics (Part C), 37(2): 258-267 (2007).

[4] McConnell, R. K., “Supervised method for optimum hyperspectral band selection”, in Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIX, Sylvia S. Shen; Paul E. Lewis, Editors, Proceedings of SPIE Vol. 8743 (SPIE, Bellingham, WA 2013).

[5] McConnell, R. K., “Train-by-show in color-based assembly and packaging inspection”, Machine Vision Applications, Architectures, and Systems Integration VI, Solomon, S. S., Batchelor, B. G., and Miller, J. W. V., editors, Proc. SPIE 3205, 210-217 (1997).